Mechanisms with Evidence: Commitment and Robustness[1]

Elchanan Ben-Porath [2] Eddie Dekel [3] Barton L. Lipman[4]

First Draft
January 2017

[1] We thank the National Science Foundation, grant SES–0820333 (Dekel), and the US–Israel Binational Science Foundation for support for this research.

[2] Department of Economics and Center for Rationality, Hebrew University. Email: benporat@math.huji.ac.il.

[3] Economics Department, Northwestern University, and School of Economics, Tel Aviv University. Email: dekel@northwestern.edu.

[4] Department of Economics, Boston University. Email: blipman@bu.edu.

Abstract

We show that in a class of I–agent mechanism design problems with evidence, commitment is unnecessary, randomization has no value, and robust incentive compatibility has no cost. In particular, for each agent i, we construct a simple disclosure game between the principal and agent i where the equilibrium strategies of the agents in these disclosure games give their equilibrium strategies in the game corresponding to the mechanism but where the principal is not committed to his response. In this equilibrium, the principal obtains the same payoff as in the optimal mechanism with commitment. As an application, we show that certain costly verification models can be characterized using equilibrium analysis of an associated model of evidence.

1 Introduction

We show that in a class of I–agent mechanism design problems with evidence, randomization has no value for the principal and robust incentive compatibility has no cost. Also, commitment is unnecessary in the sense that there is an equilibrium of the game when the principal is not committed to the mechanism with the same outcome as in the optimal mechanism with commitment. We also show that this equilibrium can be computed from a collection of I auxiliary games, where the ith game is a simple disclosure game between agent i and the principal. As an application, we show that certain mechanism design problems with costly verification instead of evidence can be solved via an associated evidence model.[1]

To understand the class of mechanism design problems we consider, consider the following examples.

Example 1. The simple allocation problem. The principal has a single unit of an indivisible good which he can allocate to one of I agents. Each agent has a type which affects the value to the principal of allocating the good to that agent. Each agent prefers getting the good to not getting it, regardless of her type. Types are independent across agents and monetary transfers are not possible. Each agent may have concrete evidence which proves to the principal some facts about her type. For example, the principal may be a dean with one job slot to allocate to a department in the College. Each department wants the slot and each has private information regarding the characteristics of the person the department would likely hire with the slot, information that is relevant to the value to the dean of assigning the slot to the department. Alternatively, the principal may be a state government which needs to choose a city in which to locate a public hospital. The state wants to place the hospital where it will be most efficiently utilized, but each city wants the hospital and has private information on local needs. The state could ask the city to bear the cost of the hospital, but that would imply diverting the city's funds from other projects that the government considers important.

Extensions: More complex allocation problems. A broader class of allocation problems will also fit in our framework. For example, consider again the example of a dean given above, but suppose the dean has several job slots to allocate where each department can have at most one and there are fewer slots than departments. A related problem is the allocation of a budget across divisions by the head of a firm. Suppose the organization has a fixed amount of money to allocate and that the value produced by a division is a function of its budget and its privately known productivity. Alternatively, consider a task allocation problem where the principal is a manager who must choose an

[1] In a model with costly verification, the agents do not have evidence to present but the principal can learn the true type of an agent at a cost.

employee to carry out a particular job. Suppose none of the employees wants to do the task and each has private information about how well he would do it. Finally, we could consider a task that some employees might want to do and others would not want to do, where both the employee's ability and desire to do the job are private information.

Example 2. The public goods problem. The principal has to choose whether or not to provide a public good which affects the utility of I agents. If the principal provides the good, the cost must be evenly divided among the agents. Each agent has a type which determines her willingness to pay for the good. If the willingness to pay exceeds her share of the cost, she wants the good to be provided and otherwise prefers that it not be provided. Types are independent across agents and monetary transfers other than the cost sharing are not possible. Each agent may have evidence which enables her to prove some facts to the principal about the value of the public good to her. The principal wishes to maximize the sum of the agents' utilities. For example, the principal may be a government agency deciding whether or not to build a hospital in a particular city and the agents may be residents of that city who will be taxed to pay for the hospital if it is built. Then an agent might show documentation of a health condition or past emergency room visits to prove to the principal that she has a high value for a nearby hospital. Alternatively, the principal can maximize a weighted sum of the agents' utilities plus a utility of her own for the public good.

We will show that optimal mechanisms for these examples share several significant features. First, commitment is not necessary. In other words, if the principal is not committed to the mechanism, there is still an equilibrium of the game with the same outcome as in the optimal mechanism. Second, the optimal mechanism is deterministic — the principal does not need to randomize. Third, the optimal mechanism is not just incentive compatible but is also what we will call *robustly incentive compatible*. We define this precisely later, but for now simply note that it is a strengthening of dominant strategy incentive compatibility. Thus the robustness of dominant strategy incentive compatibility comes at no cost to the principal. The robustness of the mechanism in turn implies similar robustness properties of the equilibrium which achieves the same outcome.[2]

One useful implication of this result is that we can compute optimal mechanisms by considering equilibria of the game without commitment. In particular, we give a relatively simple characterization of an optimal equilibrium for the principal which does not rely on much information regarding the principal's preferences or the structure of the set of actions. More specifically, we construct a collection of I auxiliary games, one for each agent, where the game for agent i is a simple disclosure game between agent

[2] This does not mean that the "truth telling" strategies used in the mechanism are also used in the game without commitment. In general, agents may be mixing over reports and evidence in the equilibrium of the game.

i and the principal. The equilibrium of the game without commitment between the I agents and the principal which has the same outcome as the optimal mechanism can be constructed by assigning to agent i her equilibrium strategy in her auxiliary game. This makes determining the optimal mechanism straightforward in some cases.

To illustrate, we consider optimal mechanisms when the evidence technology is the one originally proposed by Dye (1985). In Dye's model, each agent has some probability of having evidence that would enable her to exactly prove her true type and otherwise has no evidence at all. When we apply this approach to the simple allocation problem described in Example 1 above or to the public good problem of Example 2, we find optimal mechanisms reminiscient of optimal mechanisms in a different context, namely, under costly verification. We discuss this connection to Ben-Porath, Dekel, and Lipman (2014) and to Erlanson and Kleiner (2015) in Section 5 where we show that a class of costly verification models can be solved using our results for evidence models.

The paper is organized as follows. Section 2 presents the formal model. In Section 2.5, we state the main results sketched above, including the characterization of the best equilibrium for the principal. The proof of this theorem is sketched in Section 4. In Section 3, we specialize to the Dye (1985) evidence structure and provide a characterization of optimal mechanisms in this setting. We then use this characterization to give optimal mechanisms for a variety of more specific settings including the simple allocation problem and the public goods problem. We also show that under some conditions, optimal mechanisms for costly verification instead of evidence can be solved using the optimal mechanisms for Dye evidence. We offer concluding remarks in Section 5. Proofs not contained in the text are in the Appendix.

Related literature. Our work is related to the literature on mechanism design with evidence. The first paper on this topic was Green and Laffont (1986). We make use of results in Bull and Watson (2007) and Deneckere and Severinov (2008).[3] A particularly relevant subset of this literature is a set of papers on one–agent mechanism design problems which show that, under certain conditions, the principal does not need commitment to obtain the same outcome as under the optimal mechanism. This was first shown by Glazer and Rubinstein (2004, 2006) and extended by Sher (2011) and by Hart, Kremer, and Perry (2016, forthcoming). We discuss these papers in more detail in Section 5.

Also, our result showing that commitment is not necessary can be thought of as a characterization of equilibria in games with evidence. Hence our work is also related to the literature on communication games with evidence. The first papers on this topic are Grossman (1981) and Milgrom (1981). Our work makes particular use of Dye (1985) and Jung and Kwon (1988). Finally, the papers most closely related to our application to

[3] Other papers which are less directly related include Ben-Porath and Lipman (2012), Kartik and Tercieux (2012), and Sher and Vohra (2015).

costly verification models are Ben-Porath, Dekel, and Lipman (2014) and Erlanson and Kleiner (2015).

2 Model and Results

The set of agents is $\mathcal{I} = \{1, \dots, I\}$ where $I \geq 1$. The principal has a finite set of actions A and can randomize over these. For example, in the simple allocation problem, we have $A = I$ where $a = i$ means that the good is allocated to i. More generally, a can be interpreted as an allocation of money (where money is finitely divisible) as well as other goods, public or private. Each agent i has private information in the form of a type t_i where types are distributed independently across agents. The finite set of types of i is denoted T_i and the (full support) prior is denoted ρ_i.

2.1 Preferences

Given action a by the principal and type profile t, agent i's utility is $\bar{u}_i(a, t_i)$, independent of t_{-i}. We need significantly more structure on the agents' utility functions than this "private values" assumption, as we discuss in detail below.

The principal's utility is

$$v(a, t) = u_0(a) + \sum_i \bar{u}_i(a, t_i)\bar{v}_i(t_i).$$

For notational convenience, we define $\bar{v}_0(t_0) = 1$ so that we can write this as $\sum_i \bar{u}_i(a, t_i)\bar{v}_i(t_i)$ with the convention that the sum runs from $i = 0$ to I.

There are two ways to interpret the principal's utility function. The most obvious is a social welfare interpretation where the principal maximizes a weighted sum of the agent's utilities and $\bar{v}_i(t_i)$ determines how much he "cares" about agent i's utility. On the other hand, this utility function does not require the principal to care about the agents at all. A different interpretation is to think of $\bar{v}_i(t_i)$ as measuring the extent to which the principal's interests are aligned with those of agent i. That is, a high value of $\bar{v}_i(t_i)$ doesn't mean that the principal likes agent i but means that the principal likes what agent i likes.[4] Of course, one can also interpret the model as assuming both motivations for the principal.

[4] For example, consider the simple allocation problem where the principal is the head of an organization who needs to choose one of the agents to promote. Assume that every agent wishes to be promoted, so i's utility is 1 if he is promoted and 0 otherwise. So in this context, it is natural to assume that $\bar{v}_i(t_i)$ measures the ability of type t_i, not how much the principal cares about t_i's utility.

Another important issue for interpretation is that we cannot entirely separate assumptions about the principal's utility function and the agents' utility functions. For example, suppose $\bar{v}_i(t_i) > 0$ for all t_i and all i. Then consider changing agent i's utility function from $\bar{u}_i(a, t_i)$ to $\hat{u}_i(a, t_i) = \bar{u}_i(a, t_i)\bar{v}_i(t_i)$ and changing the principal's utility function to $\sum_i \hat{u}_i(a, t_i)$. Because $\hat{u}_i(a, t_i)$ is a positive affine transformation of $\bar{u}_i(a, t_i)$, we haven't changed best responses for the agents. Clearly, the principal's preferences have not changed since this is simply a different way of writing the same function. Hence we cannot separate $v_i(t_i)$ into the part that comes from how the principal evaluates i's utility and how agent i evaluates outcomes.

Also, note that we allow $\bar{v}_i(t_i)$ to be zero or negative. Thus the principal's interests can be in conflict with those of some or all agents in a way which depends on the agents' types.

Turning to the details of our assumptions on the agents' utility functions, we go beyond the type–independent preferences that the literature has assumed, but require that the type dependence takes a particularly simple, multiplicatively separable, form. Specifically, we say that $\bar{u}_i(a, t_i)$ satisfies *simple type dependence* if there exist functions $u_i : A \to \mathbf{R}$ and $\beta_i : T_i \to \mathbf{R}$ such that $\bar{u}_i(a, t_i) = u_i(a)\beta_i(t_i)$ where $\beta_i(t_i) \neq 0$ for all $t_i \in T_i$.[5] After explaining how these preferences capture the examples above, we provide a renormalization that gives a more useful form for analyzing the model and which also helps show how this model generalizes the cases considered in the literature.

To show that simple type dependence accommodates all the examples discussed in the introduction, we illustrate with two examples. First, consider the simple allocation problem, Example 1. Let $A = \{1, \ldots, I\}$ where $a = i$ means the principal allocates the good to agent i. Since every agent desires the good regardless of t_i, we let $\beta_i(t_i) = 1$ for all i and let $u_i(i) = 1$ and $u_i(j) = 0$ for all $j \neq i$. Finally, let $u_0(a) \equiv 0$. Then we can interpret $\bar{v}_i(t_i)$ as the value to the principal of allocating the good to agent i when his type is t_i.

As another example, consider the public goods problem, Example 2. Let $A = \{0, 1\}$, where 1 corresponds to providing the good and 0 to not providing it. Let $\beta_i(t_i)$ be the value of the public good to type t_i net of i's share of the cost of provision. Letting $u_i(a) = a$, then the utility of agent i is $\bar{u}_i(a, t_i) = u_i(a)\beta_i(t_i)$. If we take the utility of the principal to be the sum of the utilities of the agents, then letting $\bar{v}_i(t_i) = 1$ for every t_i and every i, the utility of the principal is $v(a, t) = \sum_i \bar{u}_i(a, t_i)\bar{v}_i(t_i)$.

[5] If $\beta_i(t_i) = 0$ for some t_i, then that type is indifferent over all actions by the principal and so will always truthfully reveal. Hence we may as well disregard such types.

We now renormalize the agents' utility functions. Let

$$u_i(a, t_i) = \frac{\bar{u}_i(a, t_i)}{|\beta_i(t_i)|}.$$

Clearly, $u_i(\cdot, t_i)$ represents the same preferences over $\Delta(A)$ as $\bar{u}_i(\cdot, t_i)$ for every t_i and hence the model is strategically equivalent if we use u_i for i's utility function. Note that

$$u_i(a, t_i) = \begin{cases} u_i(a), & \text{if } t_i \in T_i^+; \\ -u_i(a), & \text{if } t_i \in T_i^- \end{cases}$$

where

$$T_i^+ = \{t_i \in T_i \mid \beta_i(t_i) > 0\}$$

and $T_i^- = T_i \setminus T_i^+$. We refer to T_i^+ as the *positive types* of i and T_i^- as the *negative types* of agent i. Thus all types have indifference curves over $\Delta(A)$ defined by $u_i(a)$, though types may differ in terms of the direction of increase in utility. Also, types can differ in terms of preference intensity as measured by $\beta_i(t_i)$. This intensity factor does not have implications for i's preferences over $\Delta(A)$ but does for the principal's in the sense that a change in $\beta_i(t_i)$, all else equal, changes the principal's preferences over $\Delta(A)$ conditional on t_i.[6,7] With this rewriting of the agents' utilities, we can rewrite the principal's utility as

$$v(a, t) = \sum_i u_i(a)\beta_i(t_i)\bar{v}_i(t_i) = \sum_i u_i(a)v_i(t_i),$$

where $v_i(t_i) = \beta_i(t_i)\bar{v}_i(t_i)$ (with $\beta_0(t_0)$ defined to be 1). We will typically write the utility functions in this form henceforth.

While the assumption of simple type dependence is restrictive in general, it obviously has type–indepdendent preferences as a special case. If we set $T_i^- = \emptyset$, we have the assumption used in most of the literature on mechanism design with evidence and, in

[6] It may seem odd to have a part of the agent's utility function which is irrelevant to her preferences. We can think of β_i as measuring the intensity of i's preferences over A relative to some other actions which are not under the principal's control and do not affect the principal's choices. In other words, if agent i's utility function is $u_i(a)\beta_i(t_i) + g_i(w_i)$ where w_i is a bundle of private goods chosen by the agent, then β_i is relevant to i's preferences overall, but not for i's preferences with respect to those choices which are included in the model. For example, in the public goods problem discussed above, our interpretation of $\beta_i(t_i)$ as the monetary value of the public good to t_i minus her share of the costs implicitly treats w_i as money and assumes $g_i(w_i) = w_i$.

[7] This does not require us to assume that the principal "cares" about the intensity of the agents' preferences. If $\beta_i(t_i) > \beta_i(t_i')$, we could have $\bar{v}_i(t_i) < \bar{v}_i(t_i')$ to an extent which offsets this, leaving the principal's preferences conditional on t_i the same as his preferences conditional on t_i'. On the other hand, this formulation allows the principal to respond to differences in intensities. The public goods problem discussed above is one where the intensity of agent preferences naturally matters to the principal. For another example, consider the simple allocation problem described above where the principal is a utilitarian. Then the principal wishes to allocate the good to the agent whose preference for the good is the "most intense."

particular, the papers on the value of commitment. More broadly, in many settings, the agent has only two type–independent indifference curves over A and in this case, simple type–dependence is without loss of generality. For example, if the principal has only two actions, then, obviously, there can only be two indifference curves (at most). The type–independent version of this setting is the case originally considered by Glazer and Rubinstein (2004, 2006). Similarly, consider a type–dependent version of the simple allocation problem where each agent cares only about whether she receives the good or not, but some types prefer to get the good and others prefer not to.[8] Here the principal has as many actions as there are agents (more if she can keep the good), but each agent has only two indifference curves over A. In this case, there are only two (nontrivial) preferences over $\Delta(A)$, so this formulation is not restrictive in that context.

2.2 Evidence

Each agent may have evidence which would prove some claims about herself. To model evidence, we assume that for every i, there is a function $\mathcal{E}_i : T_i \to 2^{2^{T_i}}$. In other words, $\mathcal{E}_i(t_i)$ is a collection of subsets of T_i, interpreted as the set of events that t_i can prove. The idea is that if $e_i \in \mathcal{E}_i(t_i)$, then type t_i has some set of documents or other tangible evidence which she can present to the principal which demonstrates conclusively that her type is in the set $e_i \subset T_i$. For example, if agent i presents a house deed with her name on it, it proves that she is one of the types who owns a house. We require the following properties. First, proof is true. Formally, $e_i \in \mathcal{E}_i(t_i)$ implies $t_i \in e_i$. Second, proof is consistent in the sense that $s_i \in e_i \in \mathcal{E}(t_i)$ implies $e_i \in \mathcal{E}_i(s_i)$. In other words, if there is a piece of evidence that some type can present which does not rule out s_i, then it must be true that s_i could present that evidence. Clearly, if s_i could not present it, the evidence actually refutes the possibility of s_i. Putting these two properties together, we have $t_i \in e_i$ if and only if $e_i \in \mathcal{E}_i(t_i)$.

The last property we assume is not necessary for the model to be internally consistent but is a convenient simplifying assumption used in much of the literature. This property was introduced as the *full reports condition* by Lipman and Seppi (1995), but is more commonly referred to as *normality*, following Bull and Watson (2007). The condition says that there is one event that t_i can present which summarizes all the evidence she has available. Intuitively, this condition means that there are no time or other restrictions on the evidence an agent can present, so that she can present everything she has. Formally, the statement is that for every t_i, we have

$$\bigcap_{e_i \in \mathcal{E}_i(t_i)} e_i \in \mathcal{E}_i(t_i).$$

[8] For example, if the "good" is a task assignment as discussed in the extensions of Example 1 in the introduction, this formulation is natural.

That is, the event proved by showing all of t_i's evidence is itself an event that t_i can prove. Henceforth, we denote this maximally informative event by

$$M_i(t_i) = \bigcap_{e_i \in \mathcal{E}_i(t_i)} e_i.$$

We sometimes refer to t_i presenting $M_i(t_i)$ as presenting *maximal evidence*.

2.3 Mechanisms

Before formally defining a mechanism, we note that given our assumptions, it is without loss of generality to focus on mechanisms where the agents simultaneously make cheap talk reports of types and present evidence and where each agent truthfully reveals her type and presents maximal evidence. This version of the Revelation Principle has been shown by, among others, Bull and Watson (2007) and Deneckere and Severinov (2008). Formally, let $\mathcal{E}_i = \cup_{t_i \in T_i} \mathcal{E}_i(t_i)$ and $\mathcal{E} = \prod_i \mathcal{E}_i$. A *mechanism* is then a function $P : T \times \mathcal{E} \to \Delta(A)$.

For notational brevity, given a mechanism P, $t_i \in T_i$, $(s_i, e_i) \in T_i \times \mathcal{E}_i(t_i)$, and $(t_{-i}, e_{-i}) \in T_{-i} \times \mathcal{E}_{-i}$, let

$$\tilde{u}_i(s_i, e_i, t_{-i}, e_{-i} \mid t_i, P) = \sum_a P(a \mid s_i, e_i, t_{-i}, e_{-i}) u_i(a, t_i)$$

and

$$\hat{u}_i(s_i, e_i \mid t_i, P) = \mathrm{E}_{t_{-i}} \tilde{u}_i(s_i, e_i, t_{-i}, M_{-i}(t_{-i}) \mid t_i, P).$$

In words, $\tilde{u}_i(s_i, e_i, t_{-i}, e_{-i} \mid t_i, P)$ is agent i's expected utility under mechanism P when her type is t_i but she reports s_i, presents evidence e_i, and expects all other agents to claim types t_{-i} and report evidence e_{-i}. Then $\hat{u}_i(s_i, e_i \mid t_i, P)$ is i's expected utility from reporting (s_i, e_i) when her type is t_i and she expects the other agents to report their types truthfully and to provide maximal evidence.

A mechanism P is *incentive compatible* if for every agent i,

$$\hat{u}_i(t_i, M_i(t_i) \mid t_i, P) \geq \hat{u}_i(s_i, e_i \mid t_i, P),$$

for all $s_i, t_i \in T_i$ and all $e_i \in \mathcal{E}_i(t_i)$. In words, just as stated above, the agent finds it optimal to report her type truthfully and present maximal evidence given that every other agent does the same. The principal's expected payoff from an incentive compatible mechanism P is

$$\mathrm{E}_t \sum_a P(a \mid t, M(t)) v(a, t).$$

Our main result is that if the agents' preferences satisfy simple type dependence, then for the principal, commitment is not necessary, there is no cost to robust incentive compatibility, and randomization has no value. We now make this more precise.

Before defining our notion of robust incentive compatibility, we begin with more standard notions. A mechanism is *ex post incentive compatible* if for every agent i,

$$\tilde{u}_i(t_i, M_i(t_i), t_{-i}, M_{-i}(t_{-i}) \mid t_i, P) \geq \tilde{u}_i(s_i, e_i, t_{-i}, M_{-i}(t_{-i}) \mid t_i, P),$$

for all $s_i, t_i \in T_i$, all $t_{-i} \in T_{-i}$, and all $e_i \in \mathcal{E}_i(t_i)$. In other words, a mechanism is ex post incentive compatible if each agent i has an incentive to report honestly and present maximal evidence even if she knows all the other agents' types and that they are reporting truthfully.

Say that a reporting strategy $\sigma_j : T_j \to T_j \times \mathcal{E}_j$ is *feasible* if whenever $\sigma_j(t_j) = (s_j, e_j)$, we have $e_j \in \mathcal{E}_j(t_j)$. A mechanism is *dominant strategy incentive compatible* if for every agent i,

$$\mathrm{E}_{t_{-i}}\tilde{u}_i(t_i, M_i(t_i), \sigma_{-i}(t_{-i}) \mid t_i, P) \geq \mathrm{E}_{t_{-i}}\tilde{u}_i(s_i, e_i, \sigma_{-i}(t_{-i}) \mid t_i, P)$$

for all $s_i, t_i \in T_i$, all feasible $\sigma_{-i} : T_{-i} \to T_{-i} \times \mathcal{E}_{-i}$, and all $e_i \in \mathcal{E}_i(t_i)$. That is, a mechanism is dominant strategy incentive compatible if each agent i has an incentive to report honestly and present maximal evidence given any feasible strategies for her opponents.

In mechanisms with evidence, neither of these notions of incentive compatibility implies the other. A mechanism could be ex post incentive compatible, but an agent might want to deviate if she knew another agent were going to report (s_i, e_i) where $e_i \neq M_i(s_i)$. That is, an agent might want to deviate from truth telling and maximal evidence if she knew another agent was going to deviate from truth telling and maximal evidence in a detectable way. Similarly, a mechanism could be dominant strategy incentive compatible but an agent could wish to deviate if she knew the specific types of her opponents. The robustness notion we will use combines both the ex post and dominant strategy features of the above definitions.

We say that a mechanism is *robustly incentive compatible* if for every agent i,

$$\tilde{u}_i(t_i, M_i(t_i), t_{-i}, e_{-i} \mid t_i, P) \geq \tilde{u}_i(s_i, e_i, t_{-i}, e_{-i} \mid t_i, P),$$

for all $s_i, t_i \in T_i$, all $t_{-i} \in T_{-i}$, all $e_{-i} \in \mathcal{E}_{-i}$, and all $e_i \in \mathcal{E}_i(t_i)$. In other words, even if i knew the exact type and evidence reports of all other agents, it would be optimal to report truthfully and provide maximal evidence regardless of what those reports are. As noted above, ex post incentive compatibility and dominant strategy incentive compatibility are not equivalent in mechanisms with evidence even with independent private values. Robust incentive compatibility implies both ex post incentive compatibility and dominant strategy incentive compatibility, but is not implied by either. We give an example in Appendix A to illustrate.

If a mechanism is robustly incentive compatible, then it has several desirable properties. First, the mechanism does not rely on the principal knowing the beliefs of the agents about each other's types or strategies. Second, the outcome of the mechanism need not change if the agents report publicly and sequentially, rather than simultaneously, regardless of the order in which they report.

Obviously, robust incentive compatibility implies incentive compatibility, but the converse is not true. Hence the best robustly incentive compatible mechanism for the principal yields her a weakly lower expected payoff than the best incentive compatible mechanism, typically strictly lower. Our result states assumptions under which there is no difference — that is, the best incentive compatible mechanism for the principal is robustly incentive compatible.

We say a mechanism P is *deterministic* if for every $(t, e) \in T \times \mathcal{E}$, $P(t, e)$ is a degenerate distribution. In other words, for every report and presentation of evidence, whether or not it involves truth telling and maximal evidence, the principal chooses an $a \in A$ without randomizing. Of course, randomization is an important feature of optimal mechanisms in some settings. We will show that under our assumptions, there is an optimal mechanism which is deterministic.

2.4 Games

Finally, to state what it means that commitment is not necessary, we must define what the principal can accomplish in the absence of commitment. Without commitment, we assume that there is a game in which, just as in the revelation mechanism, agents simultaneously make type reports and present evidence, perhaps with randomization. The principal observes these choices and then chooses some allocation a, again perhaps with randomization. For clarity, we refer to this as the *game without commitment*. More formally, the set of strategies for agent i, Σ_i, is the set of functions $\sigma_i : T_i \to \Delta(T_i \times \mathcal{E}_i)$ such that $\sigma_i(s_i, e_i \mid t_i) > 0$ implies $e_i \in \mathcal{E}_i(t_i)$. That is, if agent i is type t_i and puts positive probability on providing evidence e_i, then this evidence must be feasible for t_i in the sense that $e_i \in \mathcal{E}_i(t_i)$.[9] The principal's set of feasible strategies, Σ_P, is the set of functions $\sigma_P : T \times \mathcal{E} \to \Delta(A)$. A belief by the principal is a function $\mu : T \times \mathcal{E} \to \Delta(T)$ giving the principal's beliefs about t as a function of the profile of reports and evidence presentation. For notational convenience, given $\sigma_{-i} \in \Sigma_{-i}$, $\sigma_P \in \Sigma_P$, $a \in A$, and $(s_i, e_i) \in T_i \times \mathcal{E}_i$, let

$$Q_i(a \mid s_i, e_i, \sigma_{-i}, \sigma_P) = \mathrm{E}_{t_{-i}} \sum_{(s_{-i}, e_{-i})} \sigma_P(a \mid s, e) \prod_{j \neq i} \sigma_j(s_j, e_j \mid t_j).$$

This is the probability the principal chooses allocation a given that she uses strategy σ_P, agents other than i use strategies σ_j, $j \neq i$, and agent i reports s_i and presents evidence

[9] We do not require t_i to report truthfully and do not require his claim of a type to be consistent with the evidence he presents. That is, we could have $\sigma_i(s_i, e_i \mid t_i) > 0$ even though $s_i \neq t_i$ and $e_i \notin \mathcal{E}_i(s_i)$.

e_i.

We study perfect Bayesian equilibria of this game. Our definition is the natural adaptation of Fudenberg and Tirole's (1991) definition of perfect Bayesian equilibrium for games with observed actions and independent types to allow type–dependent sets of feasible actions. See Appendix B for details.

The equilibria which will give the principal the same payoff as in the optimal mechanism will satisfy a certain robustness property that, for lack of a better phrase, we simply call *robustness*. Specifically, a perfect Bayesian equilibrium (σ, μ) is *robust* if for every i and every $t_i \in T_i$, $\sigma_i(s_i, e_i \mid t_i) > 0$ implies

$$(s_i, e_i) \in \arg \max_{s_i' \in T_i, e_i' \in \mathcal{E}_i(t_i)} \sum_{a \in A} \sigma_P(a \mid s_i', e_i', s_{-i}, e_{-i}) u_i(a, t_i),\ \forall (s_{-i}, e_{-i}) \in T_{-i} \times \mathcal{E}_{-i}.$$

In other words, $\sigma_i(t_i)$ is optimal for t_i regardless of the actions played by the other agents, given the strategy of the principal. Note that i's strategy is robust with respect to the strategies of the other agents, but not with respect to the principal's strategy.

Given a perfect Bayesian equilibrium (σ, μ), the principal's expected utility is

$$\mathrm{E}_t \sum_{(s,e) \in T \times \mathcal{E}} \sum_a \prod_i \sigma_i(s_i, e_i \mid t_i) \sigma_P(a \mid s, e) v(a, t).$$

We will show that there is a robust perfect Bayesian equilibrium of this game which gives the principal the same expected utility as the optimal mechanism. In this sense, the principal does not need the commitment assumed in characterizing the optimal mechanism.

When we show that commitment is unnecessary, we will construct an equilibrium with the same outcome as in the optimal mechanism. The equilibrium construction is particularly simple in that it can be constructed from a set of I one–agent games which do not depend on A or preferences over A.

Specifically, we define the *auxiliary game for agent i* as follows. This is a game with two players, the principal and agent i. Agent i has type set T_i. Type t_i has action set $T_i \times \mathcal{E}_i(t_i)$. The principal has action set $X \subseteq \mathbf{R}$ where X is the compact interval $[\min_j \min_{t_j \in T_j} v_j(t_j), \max_j \max_{t_j \in T_j} v_j(t_j)]$. Agent i's payoff as a function of t_i and the principal's choice of x is

$$\begin{cases} x, & \text{if } t_i \in T_i^+; \\ -x, & \text{otherwise.} \end{cases}$$

The principal's utility in this situation is $-(x - v_i(t_i))^2$. In other words, the artifical game is a persuasion game where positive types want the principal to believe that $v_i(t_i)$ is large and negative types want him to believe it is small. The structure of A and $u_i(a)$ play no role. As in the original game defined above, a strategy for agent i is a function $\sigma_i : T_i \to \Delta(T_i \times \mathcal{E}_i)$ with the property that $\sigma_i(s_i, e_i \mid t_i) > 0$ implies $e_i \in \mathcal{E}_i(t_i)$. We denote a strategy for the principal as $X_i : T_i \times \mathcal{E}_i \to X$.

2.5 Results: Commitment, Determinism, and Robust Incentive Compatibility

Our main results are stated in the following theorem.

Theorem 1. *If every u_i exhibits simple type dependence, then there is an optimal incentive compatible mechanism for the principal which is deterministic and robustly incentive compatible. In addition, there is a robust perfect Bayesian equilibrium of the game without commitment with the same outcome as in this optimal mechanism. In this equilibrium, agent i's strategy is also a perfect Bayesian equilibrium strategy in the auxiliary game for agent i.*

3 Optimal Mechanisms with Dye Evidence

3.1 Characterizing the Optimal Mechanism

In light of Theorem 1, we can compute the outcomes of optimal mechanisms by identifying the best perfect Bayesian equilibrium for the principal. In particular, we can compute these equilibria by considering the auxiliary game for each agent i. In some cases, these equilibria are very easy to characterize. In this section, we illustrate by considering optimal mechanisms with a particular evidence structure introduced by Dye (1985) and studied extensively in both the economics and accounting literatures. After characterizing optimal mechanisms with Dye evidence, we show that these results can also be used to characterize optimal mechanisms in a different setting. Specifically, in certain models without evidence but where the principal can verify the type of an agent at a cost, we show that the optimal mechanism can be computed from the optimal mechanism for an associated Dye evidence model.

We say that the model has Dye evidence if for every i, for all $t_i \in T_i$, either $\mathcal{E}_i(t_i) = \{T_i\}$ or $\mathcal{E}_i(t_i) = \{\{t_i\}, T_i\}$. In other words, any given type either has no evidence in the sense that she can only prove the trivial event T_i or has access to perfect evidence and can then choose between proving nothing (i.e., proving T_i) and proving exactly her type. Let T_i^0 denote the set of $t_i \in T_i$ with $\mathcal{E}_i(t_i) = \{T_i\}$. In what follows, we sometimes refer to types who have only trivial evidence as having no evidence and types with $\mathcal{E}_i(t_i) = \{T_i, \{t_i\}\}$ as having evidence.

A small complication in stating our results is that there is an essentially irrelevant but unavoidable multiplicity of equilibrium in our auxiliary games. To understand this, note that our auxiliary games differ in one respect from the usual persuasion games in the literature in that agent i both presents evidence and makes a cheap talk claim regarding her type in the former. Of course, if these cheap talk claims convey information, we can always permute agent i's use of these claims and the principal's interpretation of them to obtain another equilibrium.

There is also another form of multiplicity which is more standard in the literature on games with evidence. In some cases, we may have an equilibrium where the principal has the same beliefs about the agent whether she presents evidence e or evidence e'. In these cases, we can construct an equilibrium where the agent presents evidence e and another where she presents evidence e'.

Note that in both of these cases, the principal's beliefs about the agent along the equilibrium path are the same across these various equilibria. That is, if the agent is type t, the belief the principal will have about t is the same across these equilibria. With this issue in mind, we say that an equilibrium in the auxiliary game for agent i is *essentially unique* if all equilibria have the same outcome in this sense.

To be precise, given equilibria (σ_i^*, X_i^*) and $(\hat{\sigma}_i^*, \hat{X}_i^*)$ of the auxiliary game for i, we say these equilibria are *essentially equivalent* if for every $x \in X$ and every $t_i \in T_i$, we have

$$\begin{aligned}\sigma_i^* \left(\{(s_i, e_i) \in T_i \times \mathcal{E}_i(t_i) \mid X_i^*(s_i, e_i) = x\} \mid t_i\right)\\ = \hat{\sigma}_i^* \left(\left\{(s_i, e_i) \in T_i \times \mathcal{E}_i(t_i) \mid \hat{X}_i^*(s_i, e_i) = x\right\} \mid t_i\right).\end{aligned}$$

If there is an equilibrium with the property that every other equilibrium is essentially equivalent to it, we say the equilibrium is *essentially unique.*

The simplest case to consider with Dye evidence is where the utility functions are not type dependent at all. We say that the model exhibits *type–independent utility* if $u_i(a, t_i)$ is independent of t_i for all i and a. In other words, $T_i^- = \emptyset$, so $u_i(a, t_i) = u_i(a)$ for all t_i.

The following results build on well–known characterizations of equilibria in evidence games using the Dye evidence structure.

Theorem 2. *In any model with Dye evidence, for every i, there exists a unique v_i^* such that*

$$v_i^* = \mathrm{E}\left[v_i(t_i) \mid t_i \in T_i^0 \text{ or } v_i(t_i) \leq v_i^*\right].$$

If $T_i^- = \emptyset$, the essentially unique equilibrium in the auxiliary game for i is a pure strategy equilibrium where every type makes the same cheap–talk claim, say s_i^*, and only types with evidence and with $v_i(t_i) > v_i^*$ present (nontrivial) evidence. That is, type t_i sends $(s_i^*, e_i^*(t_i))$ with probability 1 where

$$e_i^*(t_i) = \begin{cases} T_i, & \text{if } t_i \in T_i^0 \text{ or } v_i(t_i) \leq v_i^*; \\ \{t_i\}, & \text{otherwise.} \end{cases}$$

To see the intuition, note first that cheap talk cannot be credible in this game since every type wants the principal to believe that v_i is large. So if i has no evidence (i.e., can only prove the trivial event T_i), then she has no ability to convey any information to the principal — she can only send an uninformative cheap talk message and prove nothing. If i can prove her type is t_i, she wants to do so only if $v_i(t_i)$ is at least as large as what the principal would believe if she showed no evidence. Thus types with evidence but lower values of $v_i(t_i)$ will pool with the types who have no evidence, leading to an expectation of $v_i(t_i)$ equal to v_i^*.

In this equilibrium, the principal's expectation of $v_i(t_i)$ will be v_i^* given a type with no evidence or with $v_i(t_i) \leq v_i^*$ and will equal the true value otherwise. More formally, let

$$\hat{v}_i(t_i) = \begin{cases} v_i^*, & \text{if } t_i \in T_i^0 \text{ or } v_i(t_i) \leq v_i^*; \\ v_i(t_i), & \text{otherwise.} \end{cases}$$

For every $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_I)$, let $\hat{p}(\cdot \mid \hat{v})$ denote any $p \in \Delta(A)$ maximizing

$$\sum_{a \in A} p(a) \left[u_0(a) + \sum_i u_i(a) \hat{v}_i \right].$$

The following is a corollary to Theorems 1 and 2.

Corollary 1. *In any model with type–independent utility and Dye evidence, there is an optimal mechanism P with $P(\cdot \mid t, M(t)) = \hat{p}(\cdot \mid \hat{v}(t))$. In other words, with or without commitment, the outcome selected by the principal when the profile of types is t is $\hat{p}(\cdot \mid \hat{v}(t))$.*

We can use Corollary 1 to give simple characterizations of optimal mechanisms in many cases of interest.

Example 3. The simple allocation problem (Example 1) with Dye evidence. In this case, $\hat{p}(i \mid t) > 0$ iff $\hat{v}_i(t_i) = \max_j \hat{v}_j(t_j)$. That is, the good is given to one of the agents with the highest $\hat{v}_j(t_j)$ or, equivalently, who is believed to have the highest v_j. We can break indifferences in a particularly simple way and recast this characterization in the form of a *favored–agent mechanism.*

More specifically, say that P is a favored–agent mechanism if there is a *threshold* $v^* \in \mathbf{R}$ and an agent i, the *favored agent*, such that the following holds. First, if no agent $j \neq i$ proves that $v_j(t_j) > v^*$, then i receives the good. Second, if some agent $j \neq i$ does prove that $v_j(t_j) > v^*$, then the good is given to the agent who proves the highest $v_j(t_j)$ (where this may be agent i).

Then a favored–agent mechanism where the favored agent is any i satisfying $v_i^* = \max_j v_j^*$ and the threshold v^* is given by v_i^* is an optimal mechanism. To see this, fix any

t. By definition, $\hat{v}_j(t_j) \geq v_j^*$ for all j. Hence if $v_i^* \geq v_j^*$ for all j, then $\hat{v}_i(t_i) \geq v_j^*$ for all j. Hence for any j such that $\mathcal{E}_j(t_j) = \{T_j\}$ or $v_j(t_j) \leq v_j^*$, we have $\hat{v}_i(t_i) \geq v_i^* \geq v_j^* = \hat{v}_j(t_j)$. So if every $j \neq i$ satisfies this, it is optimal for the principal to give the good to i. Otherwise, it is optimal for him to give it to any agent who proves the highest value.

As we discuss further below, this mechanism is reminiscent of the favored–agent mechanism discussed by Ben-Porath, Dekel, and Lipman (2014) for the allocation problem with costly verification. We now extend the simple allocation problem as discussed in the introduction.

Example 4. The multi–unit allocation problem with Dye evidence. It is not hard to extend the above analysis to the case where the principal has multiple identical units of the good to allocate. Suppose he has $K < I$ units and, for simplicity, assume he must allocate all of them. Suppose each agent can only have either 0 or 1 unit. Then the principal's action can be thought of as selecting a subset of $\{1, \ldots, I\}$ of cardinality K. The principal's utility given the set $\hat{\mathcal{I}}$ is $\sum_{i \in \hat{\mathcal{I}}} v_i(t_i)$. As before, agent i's utility is 0 if she does not get a unit and 1 if she does. In this case, it is easy to see that the principal allocates units to the K agents with the highest values of $\hat{v}_i(t_i)$ as computed above. It's not difficult to show that this can be interpreted as a kind of recursive favored–agent mechanism.[10]

Example 5. Allocating a "bad." Another setting of interest is where the principal has to choose one agent to carry out an unpleasant task (e.g., serve as department chair). It is easy to see that this problem is effectively identical to having $I - 1$ goods to allocate since not receiving the assignment is the same as receiving a good. Thus we can treat the principal's set of feasible actions as the set of subsets of $\{1, \ldots, I\}$ of cardinality $I - 1$, interpreted as the set of agents who are *not* assigned the task. The one aspect of this example that may seem odd is that the principal's utility if he assigns the task to agent i is then $\sum_{j \neq i} v_j(t_j)$. On the other hand, it is an innocuous renormalization of the principal's utility function to subtract the allocation–independent term $\sum_j v_j(t_j)$ from her utility. In this case, we see that the principal's payoff to assigning the task to agent i is $-v_i(t_i)$, so $v_i(t_i)$ is naturally interpreted as t_i's level of *in*competence in carrying out the task. One can apply the analysis of the previous example for the special case of $K = I - 1$ to characterize the optimal mechanism for this example.

While the case of type–independent utility with Dye evidence is particularly tractable,

[10] More specifically, we allocate the first unit to the agent with the highest value of v_i^* if no other agent proves a higher value and to the agent with the highest proven value otherwise. Once removing this agent and unit, we follow the same procedure for the second unit, and so on. It is easy to see that the agent with the highest value of v_i^* is the most favored agent in the sense that at least K agents must prove a value above her v_i^* for her to not get a unit. Similarly, the agent with the second–highest value of v_i^* is the second–most favored agent in the sense that at least $K - 1$ of the "lower ranked" agents must prove a value above her v_i^* for her not to get a unit, etc.

the case of simple type dependence is not much more difficult. To see the intuition, again consider the auxiliary game for i where some types wish to persuade the principal that $v_i(t_i)$ is large and other types want to convince him $v_i(t_i)$ is small. Suppose that when the agent doesn't prove her type, she makes a cheap talk claim regarding whether her type is positive (i.e., she wants the principal to think $v_i(t_i)$ is large) or negative (i.e., the reverse). Let v_i^+ denote the principal's belief about v_i if i does not prove her type but says it is positive and let v_i^- be the analog for the case where i claims her type is negative. If $v_i^+ > v_i^-$, then every positive type without evidence prefers to truthfully report that her type is positive and every negative type without evidence will honestly reveal that her type is negative since this leads to the best possible belief from i's point of view. If i is a positive type with evidence, she will want to prove her type only if $v_i(t_i) > v_i^+$, while a negative type with evidence will prove her type only if $v_i(t_i) < v_i^-$. Hence for this to be an equilibrium, we must have

$$v_i^+ = \mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^+ \cap T_i^0) \text{ or } (t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) \leq v_i^+)\right]$$

and

$$v_i^- = \mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^- \cap T_i^0) \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) \geq v_i^-)\right].$$

Suppose this gives a unique value for v_i^+ and v_i^-. If these values do not satisfy $v_i^+ \geq v_i^-$, then we can't have an equilibrium of this kind since the positive types who don't present evidence will prefer to act like the negative type and vice versa. In this case, we must pool all types. If we do have $v_i^+ \geq v_i^-$, then these strategies do give an equilibrium. In the case where $v_i^+ = v_i^-$, the cheap talk does not convey any extra information. When $v_i^+ > v_i^-$, cheap talk is useful, but there is another equilibrium as well where cheap talk is disregarded or treated as "babbling," as in all models with cheap talk.

The following lemma provides the background for the equilibrium characterization.

Lemma 1. *In any model with Dye evidence, for every i, there exists a unique triple v_i^+, v_i^-, and v_i^* such that*

$$v_i^+ = \mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^+ \cap T_i^0) \text{ or } (t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) \leq v_i^+)\right],$$

$$v_i^- = \mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^- \cap T_i^0) \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) \geq v_i^-)\right],$$

and

$$v_i^* = \mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^0) \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) \geq v_i^*) \text{ or } (t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) \leq v_i^*)\right].$$

Using these cutoffs, we can characterize the equilibria of the auxiliary game for i.

Theorem 3. *If $v_i^+ \leq v_i^-$, then there is an essentially unique equilibrium in the auxiliary game for i. In this pure strategy equilibrium, there is a fixed type $\hat{s}_i$ such that t_i reports $(\hat{s}_i, e_i^*(t_i))$ where*

$$e_i^*(t_i) = \begin{cases} T_i, & \text{if } t_i \in T_i^0 \text{ or } (t_i \in T_i^+ \text{and } v_i(t_i) \leq v_i^*) \text{ or } (t_i \in T_i^- \text{and } v_i(t_i) \geq v_i^*); \\ \{t_i\}, & \text{otherwise.} \end{cases}$$

If $v_i^+ > v_i^-$, there are two equilibria that are not essentially equivalent to one another and every other equilibrium is essentially equivalent to one of the two. The first is exactly the same strategy profile as above. In this second equilibrium, there are types $\hat{s}_i^+$ and $\hat{s}_i^-$ with $\hat{s}_i^+ \neq \hat{s}_i^-$ such that $t_i \in T_i^k$ sends $(\hat{s}_i^k, e_i^k(t_i))$, $k \in \{-,+\}$, where

$$e_i^+(t_i) = \begin{cases} T_i, & \text{if } t_i \in T_i^0 \text{ or } v_i(t_i) \leq v_i^+; \\ \{t_i\}, & \text{otherwise,} \end{cases}$$

and

$$e_i^-(t_i) = \begin{cases} T_i, & \text{if } t_i \in T_i^0 \text{ or } v_i(t_i) \geq v_i^-; \\ \{t_i\}, & \text{otherwise.} \end{cases}$$

Thus if $v_i^+ > v_i^-$, then there are (essentially) two equilibria in the auxiliary game. As the result below will show, we can always compare these equilibria for the principal and the better one is the one which separates the positive and negative types. Thus this is the equilibrium that corresponds to the optimal mechanism. With this in mind, now define $\hat{v}_i(t_i)$ as follows. If $v_i^+ > v_i^-$, we let

$$\hat{v}_i(t_i) = \begin{cases} v_i^+, & \text{if } t_i \in T_i^0 \cap T_i^+ \text{ or } t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) \leq v_i^+; \\ v_i^-, & \text{if } t_i \in T_i^0 \cap T_i^- \text{ or } t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) \geq v_i^-; \\ v_i(t_i), & \text{otherwise.} \end{cases}$$

If $v_i^+ \leq v_i^-$, let

$$\hat{v}_i(t_i) = \begin{cases} v_i(t_i), & \text{if } (t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) \geq v_i^*) \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) \leq v_i^*); \\ v_i^*, & \text{otherwise.} \end{cases} \tag{1}$$

For each $t \in T$, let $\hat{p}(\cdot \mid \hat{v})$ denote any $p \in \Delta(A)$ maximizing

$$\sum_{a \in A} p(a) \left[u_0(a) + \sum_i u_i(a, t_i) \hat{v}_i \right].$$

The following result is a corollary to Theorems 1 and 3.

Corollary 2. *In any model with simple type dependence and Dye evidence, there is an optimal mechanism P with $P(\cdot \mid t, M(t)) = \hat{p}(\cdot \mid \hat{v}(t))$. In other words, the outcome selected by the principal when the profile of types is t is $\hat{p}(\cdot \mid \hat{v}(t))$.*

The only part of this result that does not follow immediately from Theorems 1 and 3 is the claim above that when $v_i^+ > v_i^-$, the equilibrium that is better for the principal is the one that separates the positive and negative types. This equilibrium is better since it provides more information for the principal. This is shown in Appendix F.

Example 8. The public–goods problem. As an application, consider the public goods model discussed in Section 1. For simplicity, we write out the optimal mechanism only for the case where $v_i^+ > v_i^-$ for all i, but similar comments apply more generally. We know that in equilibrium, given a profile of types t, the principal's expectation of v_i will be given by $\hat{v}_i(t_i)$ defined in equation (1) above. Then the principal will provide the public good iff $\sum_i \hat{v}_i(t_i) > 0$.

Just as the analysis of Example 3 above was reminiscent of Ben-Porath, Dekel, and Lipman's (2014) analysis of allocation with costly verification, the optimal mechanism in this example is very reminiscent of the optimal mechanism under costly verification identified by Erlanson and Kleiner (2015).

3.2 Costly Verification

In Ben-Porath, Dekel, and Lipman (2014) and Erlanson and Kleiner (2015), costly verification is modeled by assuming the principal can pay a cost c_i to "check" or learn the realization of agent i's type, t_i. The agent cannot affect this verification process. By contrast, in the evidence model we consider here, the principal cannot acquire information about an agent without somehow inducing the agent to reveal it.

Despite this difference, the optimal mechanisms look very similar. Ben-Porath, Dekel, and Lipman identify an optimal mechanism in the costly–verification version of the simple allocation problem which is very similar to the optimal mechanism here. In both cases, there is a favored agent and a threshold. In both cases, if no non–favored agent "reports" above the threshold, the favored agent receives the object regardless of his report. Here, "reporting above the threshold" means to prove a value of $v_i(t_i)$ above the threshold. In Ben-Porath, Dekel, and Lipman, it means to make a cheap talk report of a type such that the type minus the checking cost is above the threshold. In both cases, if some non–favored agent "reports" above the threshold, the good goes to the agent with the highest such report. In the costly verification model, this is after checking this type.

Similarly, Erlanson and Kleiner consider the public goods model under costly verification. In their mechanism and in the optimal mechanism here, we compute "adjusted reports" for each agent i given t_i. In both cases, the adjusted report for a positive type is $\max\{v_i^+, v_i(t_i)\}$, while the adjusted report for a negative type is $\min\{v_i^-, v_i(t_i)\}$ for certain cutoffs v_i^+ and v_i^-. Just as with the allocation problem, the difference between these two scenarios is that the report is proven in the evidence model and is a cheap talk claim of a type adjusted by the verification cost in the costly verification model. In both the allocation and public goods problems, these reports are adjusted by the verification cost and then summed to determine the optimal action by the principal. Again, this includes some checking in the costly verification model.

These parallels are special cases of a more general result that certain costly verification models can be rewritten as a Dye evidence model, so that the optimal mechanism can be computed directly from our results about optimal mechanisms with evidence. In the subsequent text, we explain this for the simple allocation problem. We give the more general result and explain the connection to Erlanson and Kleiner in Appendix G.

So consider the following alternative model. Again, the principal has a single unit of an indivisible good to allocate, each agent prefers to receive the good, and $v_i(t_i)$ is the payoff to the principal to allocating the good to agent i. For simplicity, we assume here that $v_i(t_i) > 0$ for all t_i and all i and that no two types have the same value of $v_i(t_i)$. Instead of assuming that agents may have evidence to present, assume that the principal can pay a cost $c_i > 0$ to learn the type of agent i, which we refer to as *checking* i. As shown in Ben-Porath, Dekel, and Lipman, we can characterize an optimal mechanism as specifying functions $p : T \to \Delta(\{1, \ldots, I\})$ and $q_i : T \to [0, 1]$ where $p(t)$ gives the probability distribution over which agent the principal gives the good to as a function of type reports t and $q_i(t)$ gives the probability that the principal checks i given reports t. The principal's objective function then is

$$\mathrm{E}_t \left[\sum_i p_i(t) v_i(t_i) - q_i(t) c_i \right]$$

where $p(t) = (p_1(t), \ldots, p_I(t))$. The incentive compatibility constraints are

$$\hat{p}_i(t_i) \geq \hat{p}_i(t_i') - \hat{q}_i(t_i'), \ \forall t_i, t_i' \in T_i, \ \forall i$$

where $\hat{p}_i(t_i) = \mathrm{E}_{t_{-i}} p_i(t)$ and $\hat{q}_i(t_i) = \mathrm{E}_{t_{-i}} q_i(t)$. To see this, note that if type t_i reports truthfully, then whether he is checked or not, he will receive the good with expected probability $\hat{p}_i(t_i)$. On the other hand, if he misreports and claims to be type t_i', he will be checked with expected probability $\hat{q}_i(t_i')$. In this case, the principal will learn that he has lied and will not give him the good. Thus his probability of receiving the good is the same as t_i''s probability, minus the probability of being checked.

It is not hard to show that the solution is monotonic in the sense that $\hat{p}_i(t_i) \geq \hat{p}_i(t_i')$ if $v_i(t_i) \geq v_i(t_i')$. For each i, let t_i^0 be the type with the smallest value of $v_i(t_i)$. The monotonicity of the solution implies that if incentive compatibility holds for type t_i^0, then it holds for every other type of agent i. Hence we can rewrite the incentive compatibility constraints as

$$\hat{q}_i(t_i') \geq \hat{p}_i(t_i') - \hat{p}_i(t_i^0), \ \forall t_i' \in T_i, \ \forall i.$$

It is easy to see that the optimal solution must set $\hat{q}_i$ as small as possible since checking is costly. Hence $\hat{q}_i(t_i) = \hat{p}_i(t_i) - \hat{p}_i(t_i^0)$ for all t_i. We can then rewrite the objective function as

$$\sum_i \mathrm{E}_{t_i} \left[\hat{p}_i(t_i) v_i(t_i) - \hat{q}_i(t_i) c_i \right] = \sum_i \mathrm{E}_{t_i} \left[\hat{p}_i(t_i)(v_i(t_i) - c_i) + \hat{p}_i(t_i^0) c_i \right].$$

Thus we can solve the principal's problem by choosing p to maximize the above subject to the constraint that $\hat{p}_i(t_i) \geq \hat{p}_i(t_i^0)$ for all $t_i \in T_i$ and all i. Rewriting the objective function once more, we can write it as

$$\sum_i \mathrm{E}_{t_i}[\hat{p}_i(t_i)\tilde{v}_i(t_i)] = \mathrm{E}_t\left[\sum_i p_i(t)\tilde{v}_i(t_i)\right]$$

where

$$\tilde{v}_i(t_i) = \begin{cases} v_i(t_i) - c_i, & \text{if } t_i \neq t_i^0 \\ v_i(t_i^0) - c_i + \frac{c_i}{\rho_i(t_i^0)}, & \text{if } t_i = t_i^0. \end{cases}$$

(Recall that ρ_i is the principal's prior over T_i.)

Now consider the simple allocation problem with Dye evidence where the value to the principal of allocating the good to agent i is $\tilde{v}_i(t_i)$. Assume that $\mathcal{E}_i(t_i^0) = \{T_i\}$ and $\mathcal{E}_i(t_i) = \{\{t_i\}, T_i\}$ for all $t_i \neq t_i^0$. In this case, the objective function is the same as the one above. The incentive compatibility constraint is simply that no type who can prove his type wishes to imitate the type who cannot. That is, $\hat{p}_i(t_i) \geq \hat{p}_i(t_i^0)$, the same incentive compatibility constraint as in the costly verification model. Thus we can directly apply our characterization of optimal mechanisms with Dye evidence to derive the solution to this problem. It is straightforward to use this to give a characterization for the original costly verification model by "inverting" the $\tilde{v}_i$'s and writing the solution in terms of the original v_i's. In particular, we obtain the optimal mechanism identified by Ben-Porath, Dekel, and Lipman.

To see this, for each i, define the cutoffs $\tilde{v}_i^*$ from the $\tilde{v}_i$ functions the same way we defined v_i^* from the v_i functions. That is, $\tilde{v}_i^*$ is the expectation of $\tilde{v}_i$ conditional on t_i not having evidence (here being the type t_i^0) or having $\tilde{v}_i(t_i) \leq \tilde{v}_i^*$. As shown above, the optimal mechanism for this allocation problem with evidence is to select a favored agent who has $\tilde{v}_i^* \geq \tilde{v}_j^*$ for all $j \neq i$ and to set threshold $\tilde{v}_i^*$. This implies that it is optimal to give the good to i if $\tilde{v}_j(t_j) = v_j(t_j) - c_j \leq \tilde{v}_i^*$ for all $j \neq i$ and to give the good to that agent j who maximizes $v_j(t_j) - c_j$ otherwise. This is exactly the mechanism discussed by Ben-Porath, Dekel, and Lipman.

One can use this approach to characterize optimal mechanisms with costly verification for less simple allocation problems such as the extensions of Example 1 in Section 1 and for the model of Erlanson and Kleiner, as discussed in Appendix G.

4 Proof Sketch

In this section, we sketch the proof of Theorem 1. For simplicity, we sketch the proof in the context of a special case, namely, the simple allocation problem. So assume for this section that the principal has one unit of an indivisible good to allocate to some agent. All agents desire the good and the principal must give it to one of them. So $A = \{1, \ldots, I\}$ where $a = i$ means that the principal allocates the good to agent i. The utility functions of the agents are type independent with

$$u_i(a) = \begin{cases} 1, & \text{if } a = i \\ 0, & \text{otherwise.} \end{cases}$$

The payoff to the principal to allocating the good to agent i given type profile t is $v_i(t_i)$ which we assume is strictly positive for all i and all t_i.

One convenient simplification in the type independent case is that we can write a mechanism as a function only from type reports into choices by the principal, where it is understood that if i claims type t_i, she also reports maximal evidence for t_i, $M_i(t_i)$. This works in the type independent case because if i claims type t_i but does not show evidence $M_i(t_i)$, the principal knows how to punish i — namely, he can give i the good with zero probability, the worst possible outcome for i. This will deter any such "obvious" deviation. (Of course, the mechanism must still deter the more subtle deviations to reporting some $s_i \neq t_i$ and providing evidence $M_i(s_i)$.) So for this proof sketch, we will write a mechanism as a function $P : T \to \Delta(A)$.

Fix an optimal mechanism P. Given this mechanism, we can construct the probability that any given type of any given agent receives the good. Let

$$\hat{p}_i(t_i) = \mathrm{E}_{t_{-i}} P(i \mid t_i, t_{-i}).$$

This is type t_i's probability of being allocated the good in mechanism P. Partition each T_i according to equality under $\hat{p}_i$. In other words, for each $\alpha \in [0, 1]$, let

$$T_i^\alpha = \{t_i \in T_i \mid \hat{p}_i(t_i) = \alpha\}.$$

Of course, since T_i is finite, there are only finitely many values of α such that $T_i^\alpha \neq \emptyset$. Unless stated otherwise, any reference below to a T_i^α set assumes that this set is nonempty. Let $\mathcal{T}_i$ denote the partition of T_i so defined and $\mathcal{T}$ the induced (product) partition of T. We refer to $\mathcal{T}$ as the *mechanism partition*.

It is easy to see that incentive compatibility is equivalent to the statement that $M_i(s_i) \in \mathcal{E}_i(t_i)$ implies $\hat{p}_i(t_i) \geq \hat{p}_i(s_i)$. In other words, if t_i can report s_i credibly in the sense that t_i has available the maximal evidence of s_i, then the mechanism must give the good to t_i at least as often as s_i.

The first key observation is that without loss of generality, we can take the mechanism to be measurable with respect to the mechanism partition $\mathcal{T}$. While this property may seem technical, it is the key property behind our results and is not generally true for models with more general type dependence than we allow in Theorem 1.

To see why this property holds in the simple allocation problem, suppose it is violated. In other words, suppose we have some pair of types $s_i, s_i' \in T_i$ such that $\hat{p}_i(s_i) = \hat{p}_i(s_i')$ but P is not measurable with respect to $\{s_i, s_i'\}$ in the sense that there is some $t_{-i} \in T_{-i}$ with $P(\cdot \mid s_i, t_{-i}) \neq P(\cdot \mid s_i', t_{-i})$. Consider the alternative mechanism P^* which is identical to P unless i's report is either s_i or s_i'. For either of these actions by i, P^* specifies the *expected* allocation generated by P. In other words, if q is the probability of type s_i conditional on $\{s_i, s_i'\}$, then for every $a \in A$ and every $t_{-i} \in T_{-i}$, we set

$$P^*(a \mid s_i, t_{-i}) = P^*(a \mid s_i', t_{-i}) = qP(a \mid s_i, t_{-i}) + (1-q)P(a \mid s_i', t_{-i}).$$

It is easy to see that the incentives of agents $j \neq i$ are completely unaffected. By the private–values assumption, the payoffs to these agents don't depend on i's type directly — they are only affected by i's type through its effect on the outcome chosen by the principal. Since this change in the mechanism preserves the probability distribution over outcomes from the point of view of these agents, their incentives are unaffected.

Also, the incentives of agent i are not affected. The payoff to i from reporting anything other than s_i or s_i' are not changed. The expected payoff to i from reporting s_i was $\hat{p}_i(s_i)$ in the original mechanism, while the expected payoff from reporting s_i' was $\hat{p}_i(s_i')$. In the new mechanism, we have "averaged" these two types together, so that in the new mechanism, the probability i receives the good if she reports s_i is now $q\hat{p}_i(s_i)+(1-q)\hat{p}_i(s_i')$. But since $\hat{p}_i(s_i) = \hat{p}_i(s_i')$, this means that the probability that i receives the good if she reports s_i does not change and similarly for s_i'. Hence the expected payoff to i from every action is the same under P and P^*, so P^* must be incentive compatible.[11]

To see that this change does not affect the principal's payoff, recall that the principal's utility function is

$$v(a,t) = \sum_j u_j(a)v_j(t_j).$$

Under the original mechanism, the principal's expected payoff is

$$\begin{aligned} \mathrm{E}_t \sum_a P(a \mid t) \sum_j u_j(a)v_j(t_j) &= \sum_j \mathrm{E}_{t_j} \left[\mathrm{E}_{t_{-j}} \sum_a P(a \mid t)u_j(a) \right] v_j(t_j) \\ &= \sum_j \mathrm{E}_{t_j} \hat{p}_j(t_j)v_j(t_j). \end{aligned}$$

As noted above, for every j and every type $t_j \in T_j$, the probability t_j receives the good is unchanged in the new mechanism. Hence the expected payoff of the principal is

[11] Note that this would be true more generally if all types have the same indifference curves. In this case, if s_i is indifferent between reporting s_i or lying and claiming to be type s_i', s_i' would also be indifferent between these reports. Hence both types would have their payoffs unchanged if we replace the response to either report with the averaged response. This is the key way we use our assumption of simple type dependence.

unchanged as well. So without loss of generality, we can change from P to P^*. Repeating as needed, we construct an alternative optimal mechanism which is measurable with respect to $\mathcal{T}$.

To see why this property is critical, fix any event $\hat{T}$ in the mechanism partition $\mathcal{T}$ and suppose that the principal learns *only* that the type profile t is contained in $\hat{T}$. Since the mechanism is measurable with respect to $\mathcal{T}$, it specifies the same response for every $t \in \hat{T}$. Suppose, though, that this response is not sequentially rational for the principal in the sense that learning that $t \in \hat{T}$ would lead him to strictly prefer some different response from what the mechanism specifies so that the commitment to the mechanism is crucial. If this is true, we can consider the following alternative mechanism. If $t \notin \hat{T}$, the new mechanism is the same as the original one. If $t \in \hat{T}$, then the new mechanism chooses the same allocation as the old one with probability $1 - \varepsilon$ and chooses the strictly better response for the principal with probability ε for some sufficiently small $\varepsilon > 0$. It is easy to see that this alternative mechanism must give the principal a strictly higher expected payoff than the original mechanism. Hence if the new mechanism is incentive compatible, we have a contradiction to the optimality of the original mechanism.

To see that the new mechanism is incentive compatible for sufficiently small $\varepsilon > 0$, fix any t_i and any $s_i \neq t_i$ with $M_i(s_i) \in \mathcal{E}_i(t_i)$. Since the original mechanism is incentive compatible, we know that $\hat{p}_i(t_i) \geq \hat{p}_i(s_i)$. Suppose in the original mechanism, we have $\hat{p}_i(t_i) = \hat{p}_i(s_i)$. Since the new mechanism is measurable with respect to $\mathcal{T}$, we must have this same equality in the new mechanism. Hence in this case, t_i has no incentive to claim to be s_i. So suppose in the original mechanism, we had $\hat{p}_i(t_i) > \hat{p}_i(s_i)$. Then if we choose ε sufficiently small, the interim probabilities must still satisfy this condition. Hence the new mechanism must be incentive compatible, a contradiction.

This result is the key to Theorem 1. This argument shows that if the principal learns *only* the event of the mechanism partition containing t, then he wants to follow the mechanism — commitment would not be needed. To complete the proof, we need to show that we can construct an equilibrium of the game without commitment where this is all the useful information the principal receives. This will imply that the principal is sequentially rational, at least on the equilibrium path. As we explain below, we construct an equilibrium where the principal receives at least this information and does not receive more information that he can use.

The result above also tells us a great deal about the structure of the optimal mechanism. In particular, for any event $\hat{T}_i$ in the partition of T_i, let $\bar{v}_i(\hat{T}_i) = \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i]$. Given any event $\hat{T} = \prod_j \hat{T}_j$, the optimal mechanism must give the good to some agent $i \in \arg\max_j \bar{v}_j(\hat{T}_j)$. Ignoring ties to keep the discussion simple, we see that this pins down the optimal mechanism given the partition $\mathcal{T}$ and that the optimal mechanism is deterministic.[12] In particular, we can take the optimal mechanism to be measurable with respect to these beliefs.

This result also implies that the mechanism is robustly incentive compatible. Because the mechanism allocates the good to that agent i with the highest estimated v_i, incentive compatibility implies that every agent i must maximize $\bar{v}_i(\hat{T}_i)$ by her honest report. Hence whatever i thinks the other agents are doing, she could not do better by misreporting.

Intuitively, this is also why our auxiliary game construction works. In the mechanism, each agent i's report is chosen to try to persuade the principal that v_i is large. Strategically, then, we can analyze i's incentives by studying a persuasion game with only agent i where this is i's only objective.

We conclude this proof sketch by explaining how we use the auxiliary games to construct equilibrium strategies which obtain the same outcome in the game without commitment as in the optimal mechanism.

This construction has four steps. First, we consider equilibria in what we call the *restricted auxiliary game for i*. In this game, type t_i is restricted to sending evidence she has which is maximal evidence for some s_i which is in the same event of the mechanism partition as t_i. That is, if $s_i, t_i \in T_i^\alpha$ for some α, then in the restricted auxiliary game for i, t_i can send evidence $M_i(s_i)$ if $M_i(s_i) \in \mathcal{E}_i(t_i)$. Agent i's objective in this game is simply to make the principal believe that $v_i(t_i)$ is as high as possible. In other words, we let the principal's action in this game be the choice of a number x where his payoff is $-(x - v_i(t_i))^2$ and where agent i's utility is x. Thus x is, in effect, the principal's estimate of $v_i(t_i)$. In this game, the principal *must* learn at least that $t_i \in T_i^\alpha$ since, by construction, the only messages available to t_i reveal that $t_i \in T_i^\alpha$.

Second, we show the claim mentioned above: that given this information, the principal cannot do better than to implement the outcome of the optimal mechanism. More specifically, for each i, fix an equilibrium of the restricted auxiliary game. For any $t_i \in T_i$, the equilibrium strategy followed by t_i in the restricted game for i determines the principal's equilibrium expected value of v_i which we denote $v_i^*(t_i)$.[13] For a profile of types t, let $v^*(t) = (v_1^*(t_1), \ldots, v_I^*(t_I))$. Then the second step is to show that given type profile t, it is optimal for the principal to follow the allocation prescribed by the optimal mechanism given t when his expectation of the v_i's is given by $v^*(t)$. In this sense, the equilibrium does not give him information which he can use to improve on the

[12] The result also makes clear that even when there are ties, we can change the mechanism to be deterministic without affecting the principal's payoff. In the proof, we show how this can be done without affecting the agents' incentives.

[13] Agent i's equilibrium strategy in the restricted game could be mixed, but optimality then requires that the principal's belief in response to every pure strategy in the support must be the same. Hence the principal's belief in response to the equilibrium strategy of any type is unambiguously defined.

mechanism.

To see why this holds, suppose there is some t' for which the principal strictly prefers a different allocation of the good given beliefs $v^*(t')$ than what the mechanism calls for. We derive a contradiction similarly to the way we showed the sequential rationality property above. That is, let $\hat{T} = \prod_i \hat{T}_i$ be the event of the mechanism partition containing t'. Consider the following alternative mechanism. If $t \notin \hat{T}$, the new mechanism is the same as the original one. If $t \in \hat{T}$, then the new mechanism chooses the same allocation as the old one with probability $1 - \varepsilon$. With probability ε, the new mechanism chooses any optimal allocation given the beliefs $v^*(t)$ which is measurable with respect to these beliefs. By assumption, this new mechanism yields the principal a strictly higher payoff for any $\varepsilon > 0$. Hence this mechanism cannot be incentive compatible.

But the new mechanism is incentive compatible. Fix any $s_i, t_i \in T_i$ with $M_i(s_i) \in \mathcal{E}_i(t_i)$. If s_i is a type that t_i strictly prefers in the original mechanism not to imitate, then for ε sufficiently small, t_i prefers not to imitate s_i in the new mechanism. Hence if there is a violation of incentive compatibility, s_i and t_i must be in the same event of the mechanism partition and hence must both be in $\hat{T}_i$, the only event of the mechanism partition where the new mechanism differs from the old one. But if t_i prefers imitating s_i, it must be because s_i gets the good strictly more often than t_i. Given our construction, this must mean that $v_i^*(s_i) > v_i^*(t_i)$. But given that these beliefs come from an equilibrium of the restricted auxiliary game, this means that t_i cannot play s_i's equilibrium strategy and hence $M_i(s_i) \notin \mathcal{E}_i(t_i)$. Hence this does not give a violation of incentive compatibility.

Before turning to the third step, we show a useful implication. Fix events T_i^α and T_i^β from the partition $\mathcal{T}_i$ where $\alpha > \beta$. (Recall that T_i^α is the set of t_i who receive the good with probability α in the optimal mechanism.) Then $v_i^*(t_i) \geq v_i^*(t_i')$ for all $t_i \in T_i^\alpha$ and $t_i' \in T_i^\beta$. To see this, note that the argument above shows that the probability that $v_i^*(t_i) = \max_j v_j^*(t_j)$ must be at least α, while the probability that $v_i^*(t_i') > \max_{j \neq i} v_j^*(t_j)$ cannot exceed β. Since $\alpha > \beta$, this requires $v_i^*(t_i) \geq v_i^*(t_i')$.

The third step is to use this fact to show that by appropriately specifying beliefs in response to evidence which is not maximal for any type, we obtain an equilibrium of the *unrestricted auxiliary game for i*, where the payoffs are the same as in the restricted game but where t_i can send any evidence she possesses. Specifically, in the unrestricted auxiliary game for i, if i presents evidence e_i which is not maximal for any type, then the principal puts probability 1 on any type t_i consistent with e_i which minimizes $v_i(t_i)$.

To see why this works, consider a deviation by type t_i in the unrestricted auxiliary game to a message that was not available to her in the restricted game. By construction, this is either evidence which is not maximal for any type or is maximal evidence but for a type in a different event of the mechanism partition. First, consider the case of a deviation to evidence which is not maximal for any type. Since t_i could present $M_i(t_i)$

in the restricted game, the belief in response to her equilibrium strategy in the restricted game must be at least as large as the belief in response to $M_i(t_i)$. The belief in response to $M_i(t_i)$ must be at least $v_i(s_i)$ for any s_i which minimizes $v_i(s_i)$ subject to $M_i(t_i) \in \mathcal{E}_i(s_i)$. Since $M_i(t_i)$ rules out as many s_i's as t_i can rule out, this must be weakly larger than the belief in response to any evidence t_i can send that is not maximal for any type. Hence t_i would not deviate to such evidence.

So consider a deviation to evidence which is maximal for some other type but where this type is in a different event of the mechanism partition. That is, if t_i is in partition event T_i^α, then the deviation is to $M_i(s_i)$ for s_i in some other partition event T_i^β for $\beta \neq \alpha$. By incentive compatibility, the fact that t_i can send s_i's maximal evidence implies $\beta < \alpha$. Given the result above, then, the belief t_i generates in equilibrium must be greater than or equal to the belief s_i generates, so t_i has no incentive to deviate to s_i's equilibrium strategy. Hence t_i has no incentive to deviate to any evidence s_i could feasibly send, including $M_i(s_i)$.

In the fourth and final step, we construct an equilibrium of the game without commitment by taking the strategies of the agents and the beliefs of the principal to be the same as in the unrestricted auxiliary games and the strategy of the principal to be optimal given his beliefs and measurable with respect to them. Since the strategy of the principal is optimal given his beliefs and since his beliefs are consistent with the strategies of the agents, this is an equilibrium of the game without commitment as long as no agent wishes to deviate. But it is easy to see that an agent can gain only by deviating to a strategy which yields a larger belief about v_i. Since the strategy of agent i comes from a game where this is precisely what she maximizes, no such deviation is possible. Finally, as shown in the second step, even though the principal obtains more information in the game than the partition $\mathcal{T}$, following what the mechanism prescribes and ignoring this extra information is optimal for the principal. Hence the payoff of the principal in this equilibrium is exactly his payoff in the optimal mechanism.

5 Connection to the Literature

In this section, we discuss in more detail how our results relate to the literature. As mentioned in Section 1, there are earlier results for the one–agent setting showing that commitment is not necessary to the principal. Our result extends these in several ways. First, we consider multiple agents. Second, because we have multiple agents, we can consider robust incentive compatibility — that is, the question of robustness with respect to agents' beliefs about other agents, an issue absent in the one–agent setting. Third, our characterization of these equilibrium strategies is novel.

Even when we restrict our analysis to the case of $I = 1$ so that we also only have one agent, our results are not nested by the previous literature. Most significantly, all previous results in the literature assume that the one agent's preferences are independent of her type, while we allow simple type–dependence. So to clarify the relationship to the literature, for the remainder of this discussion, we discuss the one–agent case, so t refers to the type of the single agent, T her set of types, and u her utility function.

The first papers to show that a result of the form that commitment is not necessary in a mechanism design problem with evidence were Glazer and Rubinstein (2004, 2006). They used weaker assumptions on evidence as they did not assume normality, but they assumed that the principal only had two actions available and the agent's preference was type–independent. By contrast, as noted, in the one–agent, two–action case, our assumption of simple type–dependence imposes no restrictions on the preferences of the agent.

Sher (2011) generalizes the Glazer–Rubinstein result via a concavity assumption. He assumes type–independent utility for the agent and that the principal's utility can be written as a concave function of the agent's utility. In the one–agent version of our model, the principal's utility function is $v(a,t) = u_0(a) + v(t)u(a)$. Since this depends on a directly, not just through $u(a)$, it is not nested by (nor does it nest) Sher's assumptions, even in the type–independent version of our model.

Hart, Kremer, and Perry (2016, forthcoming) give versions of the Glazer–Rubinstein result which, like our result, assume normality of evidence. Again, they assume type–independent utility for the agent. Hart, Kremer, and Perry (forthcoming) assume that the principal *cannot* randomize. In addition, they weaken Sher's concavity assumption and instead assume that for each $t \in T$, the utility function of the principal over A can be written as $v(a,t) = \varphi_t(u(a))$ where given any $\mu \in \Delta(T)$, $\sum_t \mu(t)\varphi_t$ is single–peaked or, equivalently, strictly quasi–concave. Because we allow the principal's utility to depend on a directly, not only through $u(a)$, our model violates this assumption, even in the type–independent version of our model. Also, we prove that the principal does not need to randomize, while Hart, Kremer, and Perry *assume* he cannot.

Hart, Kremer, and Perry (2016) allow the principal to randomize. Their main assumption is called PUB or Principal's Uniform Best. This states that if we fix any indifference curve for the agent, then there is a point on that indifference curve which is best for the principal *independently* of t. In the one–agent version of our model, we have $v(a,t) = u_0(a) + u(a)v(t)$. Hence holding fixed the agent's utility, for any t, the best lottery over a is any p on the indifference curve which maximizes $\sum_a p(a)u_0(a)$. Thus except for the type–dependence we allow, our assumptions are nested in their one–agent case.

An additional contribution of Hart, Kremer, and Perry (forthcoming) is that they identify a refinement of equilibrium in the disclosure game that corresponds to the principal's best equilibrium. Our result showing that the principal's best equilibrium in the game without commitment can be found using I separate one–agent disclosure games is analogous in that it also provides a means to understand or compute this equilibrium. We note that Hart, Kremer, and Perry's approach is defined only for games with type–independent preferences (where we can restrict attention to messages that contain maximal evidence) and so does not appear to carry over to the more general case we consider.

Appendix

A Example

In this subsection, we give an example to show that in games with evidence, even with independent private values, the requirements of robust incentive compatibility, dominant strategy incentive compatibility, and ex post incentive compatibility differ.

The reason that ex post incentive compatibility and robust incentive compatibility are not equivalent is that robust incentive compatibility requires truth–telling to be optimal even when other agents deviate from truth–telling with maximal evidence. In the absence of evidence, the fact that we have independent private values means that agent i is unaffected by whether the claims of other agents are true or not. Hence these two notions would be the same in that case. But with evidence, we can have reports by the other agents that would be impossible under truth–telling with maximal evidence.

The reason that dominant strategy incentive compatibility is not the same as robust incentive compatibility is that dominant strategy incentive compatibility only requires that truth telling and maximal evidence be a best reply to any strategy function by the other agents. In the absence of evidence, the other agents could be playing constant strategies, implying that truth telling and maximal evidence must be a best reply to any reports by the other agents. In mechanisms with evidence, however, constant strategies may not be possible.

To see both points in a simple example, suppose $I = 2$ and $T_i = \{\alpha_i, \beta_i\}$, $i = 1, 2$. Suppose $\mathcal{E}_i(\alpha_i) = \{\{\alpha_i\}\}$ and $\mathcal{E}_i(\beta_i) = \{\{\beta_i\}\}$, $i = 1, 2$. Suppose the principal has just two actions, denoted 0 and 1. Assume $u_1(a) = a$ and $u_2(a) = 0$ for all a.[14] Say that agent i reports consistently if she reports $(\alpha_i, \{\alpha_i\})$ or $(\beta_i, \{\beta_i\})$ and reports inconsistently otherwise. Note that all three versions of incentive compatibility say that consistent reports are optimal and differ only in the circumstances under which consistent reports are required to be optimal. Assume that the prior probability that $t_2 = \beta_2$ is strictly below 1/2.

Consider the mechanism where the principal chooses $a = 1$ if one of the following is true. First, 1's report is consistent and 2's report (consistent or not) has evidence $\{\alpha_2\}$. Second, both reports are consistent and 2's evidence presentation is $\{\beta_2\}$. Third, both reports are inconsistent and 2's evidence presentation is $\{\beta_2\}$. If the reports do not satisfy one of these three conditions, then the principal chooses $a = 0$. Because 2 is indifferent

[14] It is not difficult to give a symmetric but more complex example where neither agent is completely indifferent.

between $a = 0$ and $a = 1$, the mechanism satisfies robust incentive compatibility for him. However, it is not robustly incentive compatible for 1. To see this, simply note that 1's best response to a report by 2 of $(\alpha_2, \{\beta_2\})$ is to be inconsistent.

On the other hand, this mechanism is both ex post incentive compatible and dominant strategy incentive compatible for 1. To see that it is ex post incentive compatible, note that if 2 is consistent, then 1's best response is always to be consistent, regardless of the type profile. To see that it is dominant strategy incentive compatible, note that for any feasible strategy for 2, 1's expected payoff to any consistent report is at least the probability that $t_2 = \alpha_2$, while the payoff to any inconsistent report is at most the probability that $t_2 = \beta_2$. Since the former strictly exceeds the latter, reporting consistently is a dominant strategy.

B Equilibrium Definition

Our definition of perfect Bayesian equilibrium is identical to that of Fudenberg and Tirole (1991) but adapted to allow type–dependent sets of feasible actions.

We say that $(\sigma_1, \ldots, \sigma_I, \sigma_P, \mu)$ is a perfect Bayesian equilibrium if the following conditions hold. First, for every i and every $t_i \in T_i$, $\sigma_i(s_i, e_i \mid t_i) > 0$ implies

$$(s_i, e_i) \in \arg \max_{s_i' \in T_i, e_i' \in \mathcal{E}_i(t_i)} \sum_{a \in A} Q_i(a \mid s_i', e_i', \sigma_{-i}, \sigma_P) u_i(a, t_i).$$

Second, for every $(s, e) \in T \times \mathcal{E}$, $\sigma_P(a \mid s, e) > 0$ implies

$$a \in \arg \max_{a' \in A} \sum_{t \in T} \mu(t \mid s, e) v(a', t).$$

Third, for every (s, e), $\mu(\cdot \mid s, e)$ respects independence across agents. That is, i's report (s_i, e_i) only affects the principal's beliefs about t_i and his beliefs about t_i and t_j respect independence for all $i \neq j$. Formally, we have functions $\mu_i : T_i \times \mathcal{E}_i \to \Delta(T_i)$ such that for all $t \in T$ and all $(s, e) \in T \times \mathcal{E}$,

$$\mu(t \mid s, e) = \prod_i \mu_i(t_i \mid s_i, e_i).$$

Fourth, for all (s, e), $\mu(\cdot \mid s, e)$ respects feasibility. That is, the principal's beliefs must put zero probability on any type which is infeasible given (s, e). Formally, for every $t_i \in T_i$ and $(s_i, e_i) \in T_i \times \mathcal{E}_i$, we have $\mu_i(t_i \mid s_i, e_i) = 0$ if $e_i \notin \mathcal{E}_i(t_i)$.

Finally, the principal's beliefs are consistent with Bayes' rule whenever possible in the sense that for every $(s_i, e_i) \in T_i \times \mathcal{E}_i$ such that there exists t_i with $\sigma_i(s_i, e_i \mid t_i) > 0$, we have

$$\mu_i(t_i \mid s_i, e_i) = \frac{\sigma_i(s_i, e_i \mid t_i)\rho_i(t_i)}{\sum_{t'_i \in T_i} \sigma_i(s_i, e_i \mid t'_i)\rho_i(t'_i)}.$$

(Recall that ρ_i is the principal's prior over t_i.)

C Proof of Theorem 1

Throughout the Appendix, we assume that each u_i satisfies simple type dependence and consequently write the agents' utility functions as

$$u_i(a_i, t_i) = \begin{cases} u_i(a), & \text{if } t_i \in T_i^+; \\ -u_i(a), & \text{if } t_i \in T_i^-, \end{cases}$$

where $T_i^+ \cup T_i^- = T_i$. We write the principal's utility function as

$$v(a,t) = u_0(a) + \sum_i u_i(a)v_i(t_i) = \sum_{i=0}^{I} u_i(a)v_i(t_i),$$

where we rewrite as explained in Section 2.1.

For each $i = 1, \ldots, I$, let $R_i \equiv T_i \times \mathcal{E}_i$. Given a mechanism P and $r_i \in R_i$, let

$$\hat{u}_i(r_i; P) = \mathrm{E}_{t_{-i}} \sum_a P(a \mid r_i, t_{-i}, M_{-i}(t_{-i}))u_i(a).$$

Recall that the Revelation Principle for this class of problems says that we can restrict attention to equilibria where t_i honestly states her type and provides maximal evidence. Hence $\hat{u}_i(r_i; P)$ will be the expected utility of t_i from report r_i in the mechanism.

Fix an optimal mechanism P. For each $\alpha \in \mathbf{R}$, let

$$R_i^\alpha = \{r_i \in R_i \mid \hat{u}_i(r_i; P) = \alpha\}.$$

Finiteness of T_i implies that $\mathcal{E}_i$ is finite. Given this, "most" R_i^α will be empty. When we refer to one of these sets below, we often take as given that it is nonempty. Note that the nonempty R_i^α's form a partition of R_i. In what follows, we refer to this partition as the *mechanism partition for* i, denoted $\{R_i^\alpha\}$ and refer to the product partition of R formed by the cells $\prod_i R_i^{\alpha_i}$ simply as the *mechanism partition*, denoted $\{\prod_i R_i^{\alpha_i}\}$. It will also be useful to define

$$T_i^\alpha = \{t_i \in T_i \mid \hat{u}_i(t_i, M_i(t_i); P) = \alpha\} = \{t_i \in T_i \mid (t_i, M_i(t_i)) \in R_i^\alpha\}.$$

Note that we could have some values of α with $\hat{u}_i(t_i, e_i; P) = \alpha$ only for $e_i \neq M_i(t_i)$ in which case $R_i^\alpha \neq \emptyset$ but $T_i^\alpha = \emptyset$.

Lemma 2. *Fix $(s_i, e_i) \in R_i^\alpha$. For any $t_i \in T_i^+$, if $(t_i, M_i(t_i)) \in R_i^\beta$ with $\alpha > \beta$, then we have $e_i \notin \mathcal{E}_i(t_i)$. For any $t_i \in T_i^-$, if $(t_i, M_i(t_i)) \in R_i^\beta$ with $\alpha < \beta$, we have $e_i \notin \mathcal{E}_i(t_i)$.*

Proof. Follows from incentive compatibility. ▌

Lemma 3. *Without loss of generality, we can assume the optimal mechanism P has the property that for all i and all α, if $R_i^\alpha \neq \emptyset$, then $T_i^\alpha \neq \emptyset$.*

Proof. Suppose the optimal mechanism does not have this property. Fix any $R_i^\alpha \neq \emptyset$ such that $T_i^\alpha = \emptyset$. By the Revelation Principle, the $r_i \in R_i^\alpha$ are not used in equilibrium since $T_i^\alpha = \emptyset$ implies they are all of the form (t_i, e_i) where $e_i \neq M_i(t_i)$. Intuitively, then, we can change the outcome from these off equilibrium reports so that they remain off equilibrium without changing the principal's payoff.

More specifically, choose any β such that $T_i^\beta \neq \emptyset$ and there does not exist $\gamma \in (\alpha, \beta) \cup (\beta, \alpha)$ with $T_i^\gamma \neq \emptyset$. It is easy to see that such a β must exist. First, there must be some β with $T_i^\beta \neq \emptyset$ since the nonempty T_i^β sets partition T_i. So we can simply choose the smallest $\beta > \alpha$ such that $T_i^\beta \neq \emptyset$ if such a β exists and the largest $\beta < \alpha$ with $T_i^\beta \neq \emptyset$ otherwise.

Fix any $(\hat{t}_i, \hat{e}_i) \in R_i^\beta$ and consider the mechanism P^* given by

$$P^*(\cdot \mid t, e) = \begin{cases} P(\cdot \mid t, e), & \text{if } (t_i, e_i) \notin R_i^\alpha; \\ P(\cdot \mid \hat{t}_i, \hat{e}_i, t_{-i}, e_{-i}), & \text{otherwise.} \end{cases}$$

Note that we have only changed the mechanism for reports by i which are in R_i^α and hence are *not* of the form $(t_i, M_i(t_i))$. Hence the incentive compatibility of P for $j \neq i$ implies incentive compatibility of P^* for $j \neq i$. Similarly, the principal's payoff from P^* is the same as his payoff from P.

To see that P^* is incentive compatible for i, fix any t_i and any (s_i, e_i) such that $e_i \in \mathcal{E}_i(t_i)$. Clearly, $\hat{u}_i(t_i, M_i(t_i); P^*) = \hat{u}_i(t_i, M_i(t_i); P)$. If $(s_i, e_i) \notin R_i^\alpha$, then $\hat{u}_i(s_i, e_i; P^*) = \hat{u}_i(s_i, e_i; P)$, so the fact that t_i prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i) in P implies the same is true for P^*.

So suppose $(s_i, e_i) \in R_i^\alpha$. In this case, $\hat{u}_i(s_i, e_i; P^*) = \hat{u}_i(\hat{t}_i, \hat{e}_i; P) = \beta$, while $\hat{u}_i(s_i, e_i; P) = \alpha$. For concreteness, suppose $\beta > \alpha$ (the case where $\beta < \alpha$ is analogous). From the way we chose β, we cannot have $\alpha \leq \hat{u}_i(t_i, M_i(t_i); P) < \beta$. So either

$$\hat{u}_i(t_i, M_i(t_i); P) < \alpha = \hat{u}_i(s_i, e_i; P) < \beta = \hat{u}_i(s_i, e_i; P^*)$$

or

$$\hat{u}_i(s_i, e_i; P) = \alpha < \beta = \hat{u}_i(s_i, e_i; P^*) \leq \hat{u}_i(t_i, M_i(t_i); P).$$

Recall that $e_i \in \mathcal{E}_i(t_i)$ by assumption. Hence in the former case, incentive compatibility implies $t_i \in T_i^-$ and therefore t_i prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i) in P^*. In the latter case, incentive compatibility implies $t_i \in T_i^+$ and therefore t_i (weakly) prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i) in P^*. Either way, P^* is incentive compatible and is also an optimal mechanism. By repeating this argument, we construct an optimal mechanism with the desired property. ▌

Lemma 4. *Without loss of generality, we can take the mechanism P to be measurable with respect to the mechanism partition for each i, $\{R_i^\alpha\}$, in the sense that if $(s_i, e_i), (t_i, e_i') \in R_i^\alpha$, then $P(\cdot \mid s_i, e_i, t_{-i}, e_{-i}) = P(\cdot \mid t_i, e_i', t_{-i}, e_{-i})$ for all $(t_{-i}, e_{-i}) \in R_{-i}$. Hence we can take P to be measurable with respect to the mechanism partition $\{\prod_i R_i^{\alpha_i}\}$ in the sense that $P(\cdot \mid s, e) = P(\cdot \mid s', e')$ if $(s, e), (s', e') \in \prod_i R_i^{\alpha_i}$.*

Proof. Fix an optimal mechanism P which is not measurable in this sense. We construct an alternative mechanism which is measurable, is incentive compatible, and has the same payoff for the principal as P. Fix any i and any α such that $R_i^\alpha \neq \emptyset$. By Lemma 3, $T_i^\alpha \neq \emptyset$.

Define a mechanism P^* by $P^*(\cdot \mid s_i, e_i, t_{-i}, e_{-i}) = P(\cdot \mid s_i, e_i, t_{-i}, e_{-i})$ if $(s_i, e_i) \notin R_i^\alpha$ and otherwise

$$P^*(a \mid s_i, e_i, t_{-i}, e_{-i}) = \mathrm{E}_{t_i}(P(a \mid t_i, M_i(t_i), t_{-i}, e_{-i}) \mid (t_i, M_i(t_i)) \in R_i^\alpha),$$

for all $a \in A$ and all $(t_{-i}, e_{-i}) \in R_{-i}$.

For any agent $j \neq i$, the expected payoff under the mechanism, both from honest reporting with maximal evidence and from any deviation, is unaffected. Hence we have incentive compatibility of P^* for all $j \neq i$ from incentive compatibility of P.

For agent i for $(s_i, e_i) \in R_i^\alpha$, we have

$$\begin{aligned}
\hat{u}_i(s_i, e_i; P^*) &= \mathrm{E}_{t_{-i}}\left[\sum_a P^*(a \mid s_i, e_i, t_{-i}, M_{-i}(t_{-i}))u_i(a)\right] \\
&= \mathrm{E}_{t_{-i}}\left[\sum_a \mathrm{E}_{t_i}[P(a \mid t_i, M_i(t_i), t_{-i}, M_{-i}(t_{-i}))) \mid (t_i, M_i(t_i)) \in R_i^\alpha]u_i(a)\right] \\
&= \mathrm{E}_{t_i}\left[\mathrm{E}_{t_{-i}}\left(\sum_a P(a \mid t_i, M_i(t_i), t_{-i}, M_{-i}(t_{-i}))u_i(a)\right) \mid (t_i, M_i(t_i)) \in R_i^\alpha)\right] \\
&= \mathrm{E}_{t_i}\left[\alpha \mid (t_i, M_i(t_i)) \in R_i^\alpha\right] \\
&= \alpha = \hat{u}_i(s_i, e_i; P).
\end{aligned}$$

So every type of agent i receives the same expected payoff under the new mechanism for every report as she did in the original mechanism. Hence the incentive compatibility of P implies incentive compatibility of P^*. It is easy to see that P^* gives the principal the same expected payoff as P. Hence P^* is an optimal mechanism as well. Iterating this construction, we construct an optimal mechanism which is measurable with respect to the mechanism partition.

To see that this implies measurability with respect to the mechanism partition, apply the argument above iteratively over i. ▮

In light of Lemma 4, we henceforth assume P is measurable with respect to the mechanism partition.

Given any partition $\mathcal{R}$ of $T \times \mathcal{E}$, we say that a mechanism $\tilde{P}$ (not necessarily the optimal mechanism) is *sequentially rational given* $\mathcal{R}$ if the following is true. First, $\tilde{P}$ is measurable with respect to $\mathcal{R}$. Second, for every event E of the partition, for every $(t, M(t)) \in E$, $\tilde{P}(\cdot \mid t, M(t))$ is some $p \in \Delta(A)$ which maximizes

$$\sum_a p(a) \mathrm{E}_t \left[v(a,t) \mid (t, M(t)) \in E \right].$$

In other words, the mechanism is optimal for the principal given the information contained in the partition $\mathcal{R}$.

Lemma 5. *P is sequentially rational given the mechanism partition.*

Proof. Suppose not. Fix any $(t, M(t))$ and let $P(\cdot \mid t, M(t)) = \bar{p}$. Let $\hat{R} = \prod_j R_j^{\alpha_j}$ denote the event of the mechanism partition containing $(t, M(t))$ and suppose

$$\sum_a \tilde{p}(a) \mathrm{E}[v(a,t) \mid (t, M(t)) \in \hat{R}] > \sum_a \bar{p}(a) \mathrm{E}[v(a,t) \mid (t, M(t)) \in \hat{R}].$$

We construct a new mechanism P^* as follows. For any $(t, e) \notin \hat{R}$, $P^*(\cdot \mid t, e) = P(\cdot \mid t, e)$. For $(t, e) \in \hat{R}$,

$$P^*(\cdot \mid t, e) = (1 - \varepsilon)\bar{p} + \varepsilon \tilde{p}$$

for some small $\varepsilon > 0$. Clearly, for any $\varepsilon \in (0, 1)$, P^* yields a strictly higher payoff for the principal than P.

We now show that that for ε sufficiently small, P^* is incentive compatible. To see this, fix any $(t_i, M_i(t_i))$ and any (s_i, e_i) with $e_i \in \mathcal{E}_i(t_i)$. Suppose that under P, t_i strictly preferred reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i). Then for ε sufficiently small, this must still be true. So suppose t_i was indifferent between $(t_i, M_i(t_i))$ and (s_i, e_i). That is, $\hat{u}_i(t_i, M_i(t_i); P) = \hat{u}_i(s_i, e_i; P)$. But then by measurability of the new mechanism P^* with respect to the mechanism partition of the original mechanism P, t_i must still be indifferent between these reports in the new mechanism, so it remains incentive compatible. This contradicts the optimality of P. ▮

In what follows, for any α such that $T_i^\alpha \neq \emptyset$, let

$$\bar{v}_i(\alpha) = \mathrm{E}[v_i(t_i) \mid (t_i, M_i(t_i)) \in R_i^\alpha].$$

The following lemma will be useful.

Lemma 6. *Let*

$$\mathcal{U} = \{(\bar{u}_0, \bar{u}_1, \ldots, \bar{u}_I) \in \mathbf{R}^{I+1} \mid \exists p \in \Delta(A) \text{ with } \sum_a p(a)u_i(a) = \bar{u}_i, \ \forall i\}.$$

Given any belief of the principal over each T_i, let $\hat{v}_i$ denote the expectation of $v_i(t_i)$ under the belief over T_i and let $\hat{v} = (1, \hat{v}_1, \ldots, \hat{v}_I)$. Let $\mathcal{U}^*(\hat{v})$ denote the set of $u \in \mathcal{U}$ maximizing $\hat{v} \cdot u$. Fix any i, v, and v' such that $v_i > v'_i$ and $v'_j = v_j$ for $j \neq i$. Fix any $u \in \mathcal{U}^*(v)$ and any $u' \in \mathcal{U}^*(v')$. Then $u_i \geq u'_i$.

Proof. This result is standard, but we include a proof for completeness. Clearly, we must have

$$v \cdot u \geq v \cdot u'$$
$$v' \cdot u' \geq v' \cdot u$$

implying

$$(v - v') \cdot (u - u') \geq 0.$$

But this is $(v_i - v'_i)(u_i - u'_i)$. Since $v_i > v'_i$, we must have $u_i \geq u'_i$. ▌

Lemma 7. *For all $\alpha > \beta$, we have $\bar{v}_i(\alpha) \geq \bar{v}_i(\beta)$. In other words, (weakly) "more valuable" sets for the principal receive higher utilities.*

Proof. Fix $\alpha > \beta$. Since $\alpha > \beta$, there must exist events $T_j^{\alpha_j}$ for $j \neq i$ such that

$$\bar{u}_i(\alpha) \equiv \sum_a P(a \mid t, M(t))u_i(a) > \sum_a P(a \mid t'_i, M_i(t'_i), t_{-i}, M_{-i}(t_{-i}))u_i(a) \equiv \bar{u}_i(\beta),$$

where t_i is an arbitrary element of T_i^α, t'_i is an arbitrary element of T_i^β, and $t_j \in T_j^{\alpha_j}$ for each $j \neq i$. Let $\hat{T}_{-i} = \prod_{j \neq i} T_j^{\alpha_j}$. Let $\bar{v}_j = \bar{v}_j(\alpha_j)$ for $j \neq i$, let $\bar{v}^\alpha$ denote the vector $(\bar{v}_i(\alpha), \bar{v}_{-i})$, and define $\bar{v}^\beta$ analogously. For each $j \neq i$, let

$$\bar{u}_j(\alpha) = \sum_a P(a \mid t, M(t))u_j(a)$$

for any $t \in T_i^\alpha \times \hat{T}_{-i}$ and define $\bar{u}_j(\beta)$ analogously using any $t \in T_i^\beta \times \hat{T}_{-i}$. Finally, let $\bar{u}^\alpha$ denote the vector $(\bar{u}_i(\alpha), \bar{u}_{-i}(\alpha))$ and define $\bar{u}^\beta$ analogously. From Lemma 5, we know that $\bar{u}^\alpha$ maximizes $\bar{v}^\alpha \cdot u$ over u that can be generated by some $p \in \Delta(A)$. Similarly, $\bar{u}^\beta$ maximizes $\bar{v}^\beta \cdot u$. Since $\bar{u}_i^\alpha > \bar{u}_i^\beta$, Lemma 6 implies that $\bar{v}_i(\alpha) \geq \bar{v}_i(\beta)$. ▌

Lemma 8. *Without loss of generality, we can take the mechanism P to be measurable with respect to $\bar{v}_i$ in the sense that if $\bar{v}_i(\alpha) = \bar{v}_i(\beta)$ and $(s_i, e_i) \in R_i^\alpha$, $(s_i', e_i') \in R_i^\beta$, then $P(\cdot \mid s_i, e_i, t_{-i}, e_{-i}) = P(\cdot \mid s_i', e_i', t_{-i}, e_{-i})$ for all $(t_{-i}, e_{-i}) \in R_{-i}$. In other words, we can take the mechanism to have the property that $\alpha \neq \beta$ implies $\bar{v}_i(\alpha) \neq \bar{v}_i(\beta)$.*

Proof. Fix an optimal mechanism p which does not satisfy this property. Fix the relevant α and let $\mathcal{A} = \{\beta \mid R_i^\beta \neq \emptyset \text{ and } \bar{v}_i(\beta) = \bar{v}_i(\alpha)\}$. By assumption, there exists at least one $\beta \neq \alpha$ with $\beta \in \mathcal{A}$.

By Lemma 7, we have that $\alpha' > \beta'$ implies $\bar{v}_i(\alpha') \geq \bar{v}_i(\beta')$. Hence for any $\gamma \notin \mathcal{A}$, either γ is strictly smaller than every $\beta \in \mathcal{A}$ or γ is strictly larger than every $\beta \in \mathcal{A}$.

Define a new mechanism P^* by setting $P^*(\cdot \mid s_i, e_i, t_{-i}, e_{-i}) = P(\cdot \mid s_i, e_i, t_{-i}, e_{-i})$ if $(s_i, e_i) \notin \cup_{\beta \in \mathcal{A}} R_i^\beta$ and otherwise,

$$P^*(a \mid s_i, e_i, t_{-i}, e_{-i}) = \mathrm{E}_{t_i}[P(a \mid t_i, M_i(t_i), t_{-i}, e_{-i}) \mid (t_i, M_i(t_i)) \in \cup_{\beta \in \mathcal{A}} R_i^\beta],$$

for all $a \in A$ and all $(t_{-i}, e_{-i}) \in R_{-i}$. We now show that P^* is incentive compatible and gives the principal the same payoff as P, establishing the claim.

To see that P^* is incentive compatible, note that the interim payoff to t_j for any feasible (s_j, e_j) for $j \neq i$ is unaffected by this change. Hence we have incentive compatibility for any $j \neq i$.

So fix any t_i and any $(s_i, e_i) \neq (t_i, M_i(t_i))$ with $e_i \in \mathcal{E}_i(t_i)$. If neither $(t_i, M_i(t_i))$ nor (s_i, e_i) is contained in $\cup_{\beta \in \mathcal{A}} R_i^\beta$, then the response to either report is unaffected, so incentive compatibility of P implies that t_i prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i). If both are contained in $\cup_{\beta \in \mathcal{A}} R_i^\beta$, then the expected payoff under P^* is the same in response to either report, so this incentive compatibility constraint holds.

So suppose $(t_i, M_i(t_i)) \in \cup_{\beta \in \mathcal{A}} R_i^\beta$ and (s_i, e_i) is not. Then $(s_i, e_i) \in R_i^\gamma$ for some γ that is either below every $\beta \in \mathcal{A}$ or above every $\beta \in \mathcal{A}$. If γ is below every $\beta \in \mathcal{A}$, then $\hat{u}_i(t_i, M_i(t_i); P^*) > \hat{u}_i(s, e_i; P^*)$ and $\hat{u}_i(t, M_i(t_i); P) > \hat{u}_i(s_i, e_i; P)$. The latter inequality and the incentive compatibility of P implies $t_i \in T_i^+$, so that the former inequality implies t_i prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i). Similarly, If γ is above every $\beta \in \mathcal{A}$, then both inequalities are strictly reversed, implying $t_i \in T_i^-$ and that t_i prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i). A similar argument holds for the case where $(s_i, e_i) \in \cup_{\beta \in \mathcal{A}} R_i^\beta$ and $(t_i, M_i(t_i))$ is not. Hence P^* is incentive compatible.

To see that P^* yields the same expected payoff to the principal as P, recall that by the Revelation Principle, the specification of P^* on messages other than those of the form $(t, M(t))$ are irrelevant to the principal's payoffs. Fix any t_{-i} and let $\bar{v}_j = \bar{v}_j(R_j^{\alpha_j})$ for the α_j with $(t_j, M_j(t_j)) \in R_j^{\alpha_j}$. For any $(t_i, M_i(t_i)) \in R_i^\beta$, $\beta \in \mathcal{A}$, let $p^\beta = P(\cdot \mid$

$t_i, M_i(t_i), t_{-i}, M_{-i}(t_{-i}))$. By Lemma 5, P is sequentially rational for the principal. Hence for every $\beta \in \mathcal{A}$, p^β maximizes

$$\sum_a p^\beta(a)[u_i(a)\bar{v}_i(\beta) + \sum_{j \neq i} u_j(a)\bar{v}_j].$$

Since $\bar{v}_i(\alpha) = \bar{v}_i(\beta)$ for all $\beta \in \mathcal{A}$, the only way we can have $p^\alpha \neq p^\beta$ is if the principal is indifferent between p^α and p^β. Obviously, then, the fact that P^* differs from P in such situations has no payoff consequences. Hence P^* yields the principal the same payoff as P. ▌

In light of this result, we henceforth assume P is measurable with respect to $\bar{v}$ in the sense defined above.

Lemma 9. *P is robustly incentive compatible.*

Proof. Suppose not. Then either there exists $t_i \in T_i^+$, $e_i \in \mathcal{E}_i(t_i)$, $s_i \in T_i$, $\bar{t}_{-i} \in T_{-i}$, and $\bar{e}_{-i} \in \mathcal{E}_{-i}$ such that

$$\sum_a P(a \mid t_i, M_i(t_i), \bar{t}_{-i}, \bar{e}_{-i})u_i(a) < \sum_a P(a \mid s_i, e_i, \bar{t}_{-i}, \bar{e}_{-i})u_i(a) \tag{2}$$

or some $t_i \in T_i^-$, $e_i \in \mathcal{E}_i(t_i)$, $s_i \in T_i$, $\bar{t}_{-i} \in T_{-i}$, and $\bar{e}_{-i} \in \mathcal{E}_{-i}$ with the opposite strict inequality. Since these cases are entirely symmetric, we consider only the former, so fix t_i, s_i, e_i, $\bar{t}_{-i}$, and $\bar{e}_{-i}$ satisfying equation (2). Assume $(s_i, e_i) \in R_i^\alpha$ and $(t_i, M_i(t_i)) \in R_i^\beta$. By measurability with respect to the mechanism partition for i, we know that $\alpha \neq \beta$. By Lemma 8, $\bar{v}_i(\alpha) \neq \bar{v}_i(\beta)$. By incentive compatibility and the fact that $t_i \in T_i^+$, we have $\alpha < \beta$ and hence by Lemma 7, $\bar{v}_i(\alpha) < \bar{v}_i(\beta)$.

As in the proof of Lemma 7, for each j, including $j = i$, let

$$\bar{u}_j^\alpha = \sum_a P(a \mid t_i', M_i(t_i'), t_{-i}', M_{-i}(t_{-i}'))u_j(a)$$

for any $(t_i', M_i(t_i') \in R_i^\alpha$ and any $(t_{-i}', M_{-i}(t_{-i}')) \in \prod_{j \neq i} R_j^{\alpha_j}$. (By Lemma 3, such t_i' and t_{-i}' must exist.) Similarly, define $\bar{u}_j^\beta$ using some t_i' with $(t_i', M_i(t_i')) \in R_i^\beta$. Finally, let $\bar{v}^\alpha$ denote the vector $(\bar{v}_i(\alpha), \bar{v}_{-i})$ and define $\bar{v}^\beta$ analogously. By Lemma 6, $\bar{v}_i(\beta) > \bar{v}_i(\alpha)$ implies $\bar{u}_i^\beta \geq \bar{u}_i^\alpha$. But $\bar{u}_i^\beta$ is the left–hand side of equation (2) and $\bar{u}_i^\alpha$ is the right–hand side, a contradiction. ▌

Lemma 10. *There exists an optimal mechanism P which is robustly incentive compatible and is deterministic in the sense that $P(a \mid t, M(t)) \in \{0, 1\}$ for all $a \in A$ and $t \in T$.*

Proof. Given an arbitrary mechanism $\tilde{P}$, let $\Pi_i(\tilde{P})$ denote the mechanism partition of R_i induced by $\tilde{P}$. Given $r_i \in R_i$, let $\pi_i(r_i \mid \tilde{P})$ denote the event of $\Pi_i(\tilde{P})$ containing r_i. By

Lemmas 4 and 8, we know that there exists an optimal mechanism P which is strictly measurable with respect to $\Pi_i(P)$ for all i in the sense that if $\mathrm{E}_{t_i}[v_i(t_i) \mid (t_i, M_i(t_i)) \in \pi_i(r_i \mid P)] = \mathrm{E}_{t_i}[v_i(t_i) \mid (t_i, M_i(t_i)) \in \pi_i(r_i' \mid P)]$, then

$$P(\cdot \mid r_i, r_{-i}) = P(\cdot \mid r_i', r_{-i}),\ \forall r_{-i}.$$

In other words, P is measurable with respect to the beliefs about the v_i's induced by the mechanism partition. Let $\mathcal{P}$ denote the set of optimal mechanisms $\tilde{P}$ such that $\tilde{P}$ is strictly measurable with respect to each $\Pi_i(\tilde{P})$ in this sense. Finally, let P denote any mechanism in $\mathcal{P}$ which is minimal in the sense that there is no $P' \in \mathcal{P}$ which is strictly measurable with respect to each $\Pi_i(P')$ and for which $\Pi_i(P')$ has weakly fewer elements than $\Pi_i(P)$ for all i, strictly fewer for some i. By finiteness of T, such a P must exist.

If P is deterministic, we are done, so suppose it is not. In light of Lemma 5, this can only occur when the principal is indifferent ex post. In other words, if $P(a^* \mid r) > 0$ for some $a^* \in A$ and some $r \in R$, then

$$\sum_a P(a \mid r)\mathrm{E}_t[v(a,t) \mid (t, M(t)) \in \pi(r \mid P)] = \mathrm{E}_t[v(a^*,t) \mid (t, M(t)) \in \pi(r \mid P)].$$

Hence there must exist a deterministic mechanism, say P^*, which is strictly measurable with respect to each $\Pi_i(P)$ which yields the same expected payoff for the principal.

We now show that P^* is incentive compatible and strictly measurable with respect to each $\Pi_i(P^*)$. To show incentive compatibility, suppose P^* is not incentive compatible. Then either there exists $t_i \in T_i^+$, $s_i \in T_i$, and $e_i \in \mathcal{E}_i(t_i)$ such that

$$\hat{u}_i(t_i, M_i(t_i); P^*) < \hat{u}_i(s_i, e_i; P^*)$$

or $t_i \in T_i^-$, $s_i \in T_i$, and $e_i \in \mathcal{E}_i(t_i)$ with the reverse strict inequality. Because P^* is measurable with respect to each $\Pi_i(P)$, this implies $\hat{u}_i(t_i, M_i(t_i); P) \neq \hat{u}_i(s_i, e_i; P)$. Hence incentive compatibility of P implies

$$\hat{u}_i(t_i, M_i(t_i); P) > \hat{u}_i(s_i, e_i; P)$$

if $t_i \in T_i^+$ and the reverse strict inequality if $t_i \in T_i^-$.

Consider the mechanism $P^\lambda \equiv \lambda P + (1-\lambda)P^*$. For every $\lambda \in [0,1]$, this mechanism has the same payoff for the principal as P. Clearly, for λ sufficiently close to 1, P^λ is incentive compatible. Let λ^* be the smallest λ such that P^λ is incentive compatible. It is easy to see that such a λ^* exists and that it satisfies

$$\lambda^*\hat{u}_i(t_i, M_i(t_i); P) + (1-\lambda^*)\hat{u}_i(t_i, M_i(t_i); P^*) = \lambda^*\hat{u}_i(s_i, e_i; P) + (1-\lambda^*)\hat{u}_i(s_i, e_i; P^*)$$

for some $t_i, s_i \in T_i$ and $e_i \in \mathcal{E}_i(t_i)$ such that $\pi_i(t_i, M_i(t_i) \mid P) \neq \pi_i(s_i, e_i \mid P)$. Hence for every i, $\Pi_i(P^{\lambda^*})$ either equals or coarsens $\Pi_i(P)$, coarsening for some i. By the same reasoning as Lemmas 4 and 8, there exists another mechanism P^{**} with $\Pi_i(P^{\lambda^*}) = \Pi_i(P^{**})$ for every i which is strictly measurable with respect to each $\Pi_i(P^{**})$ and yields the principal the same expected payoff as P^{λ^*}. This contradicts the minimality of P. Hence P^* is incentive compatible.

To see that P^* is strictly measurable with respect to each $\Pi_i(P^*)$, suppose it is not. By construction, this means that each $\Pi_i(P^*)$ is either equal to or a coarsening of $\Pi_i(P)$ and is a coarsening for some i. Again following the same reasoning as Lemmas 4 and 8, there exists another mechanism P^{**} with $\Pi_i(P^*) = \Pi_i(P^{**})$ for every i which is strictly measurable with respect to each $\Pi_i(P^*)$ and yields the principal the same expected payoff as P^*. This again contradicts the minimality of P.

Finally, the same reasoning as in the proof of Lemma 9 shows that P^* is robustly incentive compatible. ▌

We now construct an equilibrium for the game which yields the same payoff for the principal as P. In particular, the strategy for agent i in this game is the same as i's strategy in an equilibrium of the auxiliary game for i. Recall that the auxiliary game for i is a two–player game between i and the principal. i has a set of types T_i where the prior over T_i is the same as in the mechanism design problem. If i is type t_i, then her set of feasible actions is $Z_i(t_i) \equiv T_i \times \mathcal{E}_i(t_i)$. The principal's set of feasible actions is $X = [\min_j \min_{t_j \in T_j} v_j(t_j), \max_j \max_{t_j \in T_j} v_j(t_j)]$. The game is sequential. First, agent i learns her type $t_i \in T_i$. Then she chooses an action $z_i \in Z_i(t_i)$. Next, the principal observes this action and chooses $x \in X$. If i's type is t_i and the principal chooses action x, then the principal's payoff is $-(x - v_i(t_i))^2$, while i's payoff is

$$\begin{cases} x, & \text{if } t_i \in T_i^+; \\ -x, & \text{otherwise.} \end{cases}$$

Denote a strategy for i in this game by $\sigma_i(\cdot \mid t_i)$, a function from T_i to $\Delta(Z_i(t_i))$. Let the principal's belief be denoted $q_i(\cdot \mid s_i, e_i)$ where this is a function from $R_i = T_i \times \mathcal{E}_i$ to $\Delta(T_i)$. Finally, the principal's action in response to (s_i, e_i) is denoted $X_i : R_i \to X$.

We construct the relevant equilibrium of the auxiliary game for i by first considering what we will call the *restricted auxiliary game.* In the restricted game, type t_i cannot choose any action in R_i but can only choose actions in R_i^α where α is the unique α such that $t_i \in T_i^\alpha$.

Fix any i and any perfect Bayesian equilibrium $(\sigma_i^*, X_i^*, q_i^*)$ of the restricted auxiliary game for i.[15] Obviously, sequential rationality implies that $X_i^*(s_i, e_i)$ is the expectation

[15] To see that such an equilibrium must exist, consider the game where i is restricted to putting probability $\varepsilon > 0$ on each of her pure strategies. By standard results, this game has a Nash equilibrium.

of $v_i(t_i)$ given the belief q_i^* or

$$\sum_{t_i \in T_i} v_i(t_i) q_i^*(t_i \mid s_i, e_i).$$

Let $\hat{X}_i^*(t_i)$ denote the action chosen by the principal when i is type t_i. That is, $\hat{X}_i^* : T_i \to X$ and is given by

$$\hat{X}_i^*(t_i) = X_i^*(s_i, e_i), \text{ for some } (s_i, e_i) \in \operatorname{supp}(\sigma_i^*(\cdot \mid t_i)).$$

Note that the principal's optimal action is always pure and that t_i is never indifferent between two distinct actions by the principal. Hence every message in the support of t_i's mixed strategy must lead to the same response by the principal. Thus the definition above is unambiguous. Clearly, for this to be an equilibrium, it must be true that if $t_i \in T_i^+$,

$$\hat{X}_i^*(t_i) = \max_{(s_i, e_i) \in Z_i(t_i) \cap R_i^\alpha} X_i^*(s_i, e_i),$$

while for $t_i \in T_i^-$,

$$\hat{X}_i^*(t_i) = \min_{(s_i, e_i) \in Z_i(t_i) \cap R_i^\alpha} X_i^*(s_i, e_i).$$

By construction, in any equilibrium of the restricted auxiliary game for i, the principal learns at least which event of the mechanism partition for i that t_i lies in. This is true because if $t_i \in T_i^\alpha$, then t_i can only send $(s_i, e_i) \in R_i^\alpha$. Hence observing (s_i, e_i) reveals the relevant value of α. Since the optimal mechanism is measurable with respect to the mechanism partition, this means that the principal must have enough information to carry out the optimal mechanism if this is the information the agents reveal to him. On the other hand, the principal may learn more than just that $t_i \in T_i^\alpha$ in the equilibrium. The following lemma shows that this extra information, if any, cannot be useful for the principal.

Lemma 11. *For each i, fix any equilibrium of the restricted auxiliary game for i and any α_i such that $T_i^{\alpha_i} \neq \emptyset$. Then for every $t \in \prod_i T_i^{\alpha_i}$,*

$$P(\cdot \mid t, M(t)) \in \arg\max_{p \in \Delta(A)} \sum_a p(a) \sum_i u_i(a) \hat{X}_i^*(t_i).$$

In other words, given the belief formed by the principal in the equilibria at profile t, it is optimal for him to follow the optimal mechanism.

As $\varepsilon \downarrow 0$ (taking subsequences as needed), these strategies converge to a Nash equilibrium of the restricted auxiliary game by upper hemicontinuity of the Nash equilibrium correspondence. These strategies and the limiting beliefs for the principal must also be a perfect Bayesian equilibrium since the principal's limiting strategy must be optimal given his limiting belief.

Proof. Suppose not. For any $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_I)$, let $\tilde{p}(\cdot \mid \hat{v})$ denote any $p(\cdot) \in \Delta(A)$ which maximizes

$$\sum_a p(a) \sum_i u_i(a)\hat{v}_i.$$

Clearly, there exists $p(\cdot \mid t)$ with

$$\sum_a p(a \mid t) \sum_i u_i(a)\hat{X}_i^*(t_i) > \sum_a P(a \mid t_i, M_i(t_i)) \sum_i u_i(a)\hat{X}_i^*(t_i) \tag{3}$$

if and only if this holds for $p(\cdot \mid t) = \tilde{p}(\cdot \mid \hat{X}^*(t))$ where $\hat{X}^*(t) = (\hat{X}_1^*(t_1), \ldots, \hat{X}_I^*(t_I))$.

Given any $(s_i, e_i) \in R_i^{\alpha_i}$, let

$$\hat{v}_i(s_i, e_i) = \begin{cases} \hat{X}_i^*(s_i), & \text{if } e_i = M_i(s_i); \\ X_i^*(s_i, e_i), & \text{otherwise.} \end{cases}$$

Given $(s, e) \in \prod_i R_i^{\alpha_i}$, let $\hat{v}(s, e) = (\hat{v}_1(s_1, e_1), \ldots, \hat{v}_I(s_I, e_I))$. Fix a small $\varepsilon > 0$ and define a new mechanism P^* by

$$P^*(\cdot \mid s, e) = \begin{cases} \varepsilon\tilde{p}(\cdot \mid \hat{v}(s, e)) + (1 - \varepsilon)P(\cdot \mid s, e), & \text{if } (s_i, e_i) \in R_i^{\alpha_i}, \ \forall i; \\ P(\cdot \mid s, e), & \text{otherwise.} \end{cases}$$

In other words, for those types $t \in \prod_i T_i^{\alpha_i}$, we assign a convex combination of the $\tilde{p}$ that will be optimal for the principal given the belief they will induce in the restricted auxiliary games and the original mechanism, assuming they report honestly and provide maximal evidence. If they deviate from maximal evidence, we assign a convex combination of the $\tilde{p}$ optimal for the principal given the induced beliefs in the restricted auxiliary games given those deviations and the original mechanism. Finally, for all other type profiles, the mechanism is unchanged.

We now show that P^* is incentive compatible. So fix some $t_i \in T_i$ and (s_i, e_i) such that $e_i \in \mathcal{E}_i(t_i)$. If t_i strictly prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i) under P, then for ε sufficiently small, t_i still has this strict preference. So suppose that t_i is indifferent between reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i) under P, so $(t_i, M_i(t_i))$ and (s_i, e_i) are in the same event of the mechanism partition for i. Clearly, if that event is not $R_i^{\alpha_i}$, then P^* still treats these reports the same way, so t_i is still indifferent.

So assume $(t_i, M_i(t_i)), (s_i, e_i) \in R_i^{\alpha}$. The only way t_i would not be indifferent under P^* is if $X_i^*(s_i, e_i) \neq \hat{X}_i^*(t_i)$. If $t_i \in T_i^+$, we know that $\hat{X}_i^*(t_i) \geq X_i^*(s_i, e_i)$. By Lemma 6, this implies

$$\begin{aligned} &\mathrm{E}_{t_{-i}}\left[\sum_a \tilde{p}(a \mid \hat{X}_i^*(t_i), \hat{X}_{-i}^*(t_{-i}))u_i(a) \mid t_{-i} \in \prod_{j\neq i} T_j^{\alpha_j}\right] \\ &\quad \geq \mathrm{E}_{t_{-i}}\left[\sum_a \tilde{p}(a \mid X_i^*(s_i, e_i), \hat{X}_{-i}^*(t_{-i}))u_i(a) \mid t_{-i} \in \prod_{j\neq i} T_j^{\alpha_j}\right]. \end{aligned}$$

Since P is incentive compatible, this implies t_i prefers reporting maximal evidence to reporting (s_i, e_i) in P^*. A similar argument applies to $t_i \in T_i^-$. Hence P^* is incentive compatible.

But then we have a contradiction. By hypothesis, P is the optimal incentive compatible mechanism, so the fact that P^* is also incentive compatible implies that it cannot yield the principal a strictly higher payoff than P. ▌

Lemma 12. *Fix $\alpha > \beta$ such that $T_i^\alpha \neq \emptyset$ and $T_i^\beta \neq \emptyset$ and any equilibrium of the restricted auxiliary game for i. Then for every $t_i \in T_i^\alpha$ and $t_i' \in T_i^\beta$, we have $\hat{X}_i^*(t_i) \geq \hat{X}_i^*(t_i')$.*

Proof. Since $\alpha > \beta$, there exists $\hat{t}_{-i} \in T_{-i}$ such that

$$u_i^\alpha \equiv \sum_a P(a \mid t_i, M_i(t_i), \hat{t}_{-i}, M_{-i}(\hat{t}_{-i}))u_i(a) > \sum_a P(a \mid t_i', M_i(t_i'), \hat{t}_{-i}, M_{-i}(\hat{t}_{-i}))u_i(a) \equiv u_i^\beta.$$

For each $j \neq i$, let

$$u_j^\alpha = \sum_a P(a \mid t_i, M_i(t_i), \hat{t}_{-i}, M_{-i}(\hat{t}_{-i}))u_j(a),$$

and define u_j^β analogously. By Lemma 11, we know that $p^\alpha \equiv P(\cdot \mid t_i, M_i(t_i), \hat{t}_{-i}, M_{-i}(\hat{t}_{-i}))$ maximizes over $p(\cdot) \in \Delta(A)$

$$\sum_a p(a) \left[u_i(a)\hat{X}_i^*(t_i) + \sum_{j \neq i} u_j(a)\hat{X}_j^*(\hat{t}_j)\right]$$

and p^β defined analogously maximizes the analog for t_i'. Hence by Lemma 6, $u_i^\alpha > u_i^\beta$ implies $\hat{X}_i^*(t_i) \geq \hat{X}_i^*(t_i')$. ▌

We now show how we can modify an equilibrium of the restricted auxiliary game for i to construct an equilibrium of the unrestricted auxiliary game for i with the same equilibrium path. So fix any equilibrium of the restricted auxiliary game for i, say $(\sigma_i^*, X_i^*, q_i^*)$. We first show that in the unrestricted auxiliary game for i, agent i does not have a profitable deviation from these strategies to any (s_i, e_i) which has positive probability in equilibrium (i.e., with $\sigma_i^*(s_i, e_i \mid t_i) > 0$ for some t_i).

Fix any $t_i \in T_i^\alpha$. Since these strategies are an equilibrium of the restricted game, t_i does not have a profitable deviation to any $(s_i, e_i) \in R_i^\alpha$ with $e_i \in \mathcal{E}_i(t_i)$. So consider a deviation by t_i to some $(s_i, e_i) \in R_i^\beta$ for $\beta \neq \alpha$ such that (s_i, e_i) has positive probability under the equilibrium strategies. By Lemma 12, we know that a deviation to any $(s_i, e_i) \in R_i^\beta$, $\beta \neq \alpha$, which has positive probability in equilibrium must at least weakly increase the principal's belief if $\beta > \alpha$ and decrease it if $\beta < \alpha$. If $t_i \in T_i^+$, then Lemma 2 implies that for every $(s_i, e_i) \in R_i^\beta$ with $\beta > \alpha$, we have $e_i \notin \mathcal{E}_i(t_i)$. Hence t_i cannot deviate to an (s_i, e_i) which has positive probability in equilibrium and increases the principal's belief. Similarly, a negative type cannot feasibly deviate to any (s_i, e_i) which has positive probability in equilibrium and decreases the principal's belief.

So we only need to ensure that there is no profitable deviation to an (s_i, e_i) which has zero probability in the restricted game. Fix any such (s_i, e_i) and suppose $(s_i, e_i) \in R_i^\alpha$. Let $F_i = \{t_i \in T_i \mid e_i \in \mathcal{E}_i(t_i)\}$. Since we have an equilibrium of the restricted game, we know that

$$\min_{\bar{t}_i \in F_i \cap T_i^+ \cap T_i^\alpha} \hat{X}_i^*(\bar{t}_i) \geq X_i^*(s_i, e_i) \geq \max_{\bar{t}_i \in F_i \cap T_i^- \cap T_i^\alpha} \hat{X}_i^*(\bar{t}_i).$$

That is, the worst off positive type who can send evidence e_i prefers the belief she induces to deviating to (s_i, e_i) and analogously for the negative types.

Let $\hat{T}_i^+$ denote the set of $t_i \in F_i \cap T_i^+$ with $t_i \in T_i^\beta$ for some $\beta \neq \alpha$. By Lemma 2, we must have $\beta > \alpha$ for each $t_i \in \hat{T}_i^+$. If $F_i \cap T_i^+ \cap T_i^\alpha \neq \emptyset$, then by Lemma 12, we have

$$\hat{X}_i^*(t_i) \geq \min_{\bar{t}_i \in F_i \cap T_i^+ \cap T_i^\alpha} \hat{X}_i^*(\bar{t}_i) \geq X_i^*(s_i, e_i), \quad \forall t_i \in \hat{T}_i^+,$$

so no positive type can gain by deviating to (s_i, e_i). So assume $F_i \cap T_i^+ \cap T_i^\alpha = \emptyset$.

Fix $\hat{t}_i \in \hat{T}_i^+$ such that $\hat{X}_i^*(\hat{t}_i) \leq \hat{X}_i^*(t_i)$ for all $t_i \in \hat{T}_i^+$. By the same reasoning as above, if $\hat{X}_i^*(\hat{t}_i) \geq X_i^*(s_i, e_i)$, no positive type can gain by deviating to (s_i, e_i). So assume $X_i^*(s_i, e_i) > \hat{X}_i^*(\hat{t}_i)$.

Suppose there is any $t_i' \in F$ such that $\hat{X}_i^*(\hat{t}_i) \geq v_i(t_i')$. If so, redefine the belief in response to (s_i, e_i) by setting it equal to a convex combination of the equilibrium belief from the restricted game and a degenerate distribution on t_i' chosen to make the expectation of v_i equal to $\hat{X}_i^*(\hat{t}_i)$. By construction, then, no positive type will be able to gain by deviating to (s_i, e_i). For any $t_i \in F_i \cap T_i^-$ with $t_i \in T_i^\gamma$, $\gamma \neq \alpha$, Lemma 2 implies $\gamma < \alpha$. Hence by Lemma 12, $\hat{X}_i^*(t_i) \leq \hat{X}_i^*(\hat{t}_i)$. Hence no negative type can gain by deviating to (s_i, e_i) either.

Hence we can assume that there is no $t_i' \in F_i$ with $\hat{X}_i^*(\hat{t}_i) \geq v_i(t_i')$. That is, every type t_i' who can send e_i has $v_i(t_i') > \hat{X}_i^*(\hat{t}_i)$. Clearly, if $\hat{t}_i$ sends $M_i(\hat{t}_i)$, an option which must be feasible in the restricted auxiliary game for i, she must prove at least as much as e_i. Hence she has a strategy available in the restricted game which must lead to a belief by the principal above $\hat{X}_i^*(\hat{t}_i)$, a contradiction.

Summarizing, we see that either no positive type can gain by sending (s_i, e_i) given the belief from the restricted auxiliary game for i this leads to or we can change that belief in such a way that no positive or negative type can gain. The symmetric argument for negative types then shows that by only changing off equilibrium beliefs, we can turn an equilibrium for the restricted auxiliary game for i into an equilibrium of the unrestricted game.

To complete the proof, we now use the equilibrium strategies of the auxiliary games to construct equilibrium strategies for the real game. The strategy for agent i is the same as her strategy in the equilibrium of the auxiliary game for i. Similarly, the principal's belief about t_i when he observes (s_i, e_i) is given by his belief in the auxiliary game for i. Given the principal's beliefs, sequential rationality tells us what his action must be at any information set where he has a unique optimal choice given his beliefs. However, we need to specify his actions at information sets with multiple optimal choices. On the equilibrium path, we will specify his actions to follow the optimal mechanism, but information sets off the equilibrium path are more subtle.

To construct the principal's equilibrium strategy, we divide the possible (s, e) profiles he may observe into three sets. First, if (s, e) has positive probability under the equilibrium strategies of the agents, the principal chooses what the mechanism prescribes given the types. To be more specific, if $(s_i, e_i) \in R_i^{\alpha_i}$ and $\sigma_i^*(s_i, e_i \mid t_i) > 0$ for some t_i for each i, then the principal chooses $P(\cdot \mid \hat{t}, M(\hat{t}))$ for any $\hat{t}$ such that $\hat{t}_i \in T_i^{\alpha_i}$ for all i. Second, if (s, e) has the property that (s_i, e_i) has zero probability under the equilibrium strategies of the agents for at least two i, then the principal chooses any optimal $p(\cdot)$ given his beliefs. Obviously, the specification of the principal's strategy on such histories does not affect equilibrium considerations for the agents.

Third, consider any (s, e) such that (s_i, e_i) has zero probability under the equilibrium strategies for exactly one i. If $X_i^*(s_i, e_i) \neq \hat{X}_i^*(t_i)$ for all t_i, then we can take the principal's response to (s, e) to be any optimal $p(\cdot) \in \Delta(A)$ given his beliefs. If $(s_i, e_i) \in R_i^\alpha$ and there exists $t_i \in T_i^\alpha$ with $X_i^*(s_i, e_i) = \hat{X}_i^*(t_i)$, then we can treat (s_i, e_i) the same way as any positive probability $(s_i', e_i') \in R_i^\alpha$ as specified above. Next, suppose $(s_i, e_i) \in R_i^\alpha$ but the only t_i's satisfying $X_i^*(s_i, e_i) = \hat{X}_i^*(t_i)$ have $t_i \notin T_i^\alpha$. If all such t_i are in the same T_i^β, then we treat (s_i, e_i) the same way as any positive probability (s_i', e_i') in R_i^β.

Finally, suppose $(s_i, e_i) \in R_i^\alpha$, there is no $t_i \in T_i^\alpha$ with $X_i^*(s_i, e_i) = \hat{X}_i^*(t_i)$, and there exists $\bar{t}_i^k \in T_i^{\beta_k}$ with $X_i^*(s_i, e_i) = \hat{X}_i^*(\bar{t}_i^k)$, $k = 1, 2$, with $\beta_1 > \beta_2$. By Lemma 12, $\beta_1 > \beta_2$ implies that every expectation of v_i induced in equilibrium by a type in $T_i^{\beta_1}$ must weakly exceed every expectation induced by a type in $T_i^{\beta_2}$. Hence it must be true that $\hat{X}_i^*(\bar{t}_i^1) = \min_{t_i \in T_i^{\beta_1}} \hat{X}_i^*(t_1)$ and $\hat{X}_i^*(\bar{t}_i^2) = \max_{t_i \in T_i^{\beta_1}} \hat{X}_i^*(t_i)$. If $\beta_2 > \alpha$, then take the principal's response to (s_i, e_i) to be the same as his response to any $(s_i', e_i') \in R_i^{\beta_2}$ which has positive probability. If $\alpha > \beta_1$, then take the principal's response to (s_i, e_i) to be the same as his response to any $(s_i', e_i') \in R_i^{\beta_1}$ which has positive probability. Finally, if $\beta_1 > \alpha > \beta_2$, take the principal's response to be a 50–50 mixture between his response given any $(s_i', e_i') \in R_i^{\beta_1}$ on the equilibrium path and the response given any $(s_i'', e_i'') \in R_i^{\beta_2}$ on the equilibrium path. (This case can only arise if $\hat{X}_i^*(\bar{t}_i^1) = \hat{X}_i(\bar{t}_i^2) = \hat{X}_i^*(t_i')$ for all $t_i' \in T_i^\alpha$.)

To see that these strategies form an equilibrium, first note that Lemma 11 implies that the principal is choosing a best reply given his beliefs in response to every (s, e) which has positive probability in equilibrium. The construction above ensures that the principal is also sequentially rational in response to any (s, e) which has zero probability in equilibrium. Turning to the agents, consider any $t_i \in T_i$ and consider a deviation by t_i to some (s_i, e_i) with $\sigma_i^*(s_i, e_i \mid t_i) = 0$. If $X_i^*(s_i, e_i) \neq \hat{X}_i^*(t_i)$, then the fact that these strategies are an equilibrium of the auxiliary game for i implies that if $t_i \in T_i^+$, we have $\hat{X}_i^*(t_i) > X_i^*(s_i, e_i)$ and the reverse strict inequality if $t_i \in T_i^-$. By Lemma 6, this implies that t_i is at least weakly worse off deviating to (s_i, e_i).

So suppose $X_i^*(s_i, e_i) = \hat{X}_i^*(t_i)$. That is, suppose the principal has the same belief about v_i under the deviation as he would following equilibrium play by t_i. For concreteness, assume $t_i \in T_i^+$ — an analogous argument covers the case where $t_i \in T_i^-$. Assume that $t_i \in T_i^\alpha$. Since $e_i \in \mathcal{E}_i(t_i)$ by hypothesis, Lemma 2 implies that $(s_i, e_i) \in R_i^\beta$ for $\beta \leq \alpha$. If there is some type $t_i' \neq t_i$ who sends (s_i, e_i) with positive probability in equilibrium, then the outcome is the same as in the optimal mechanism given any $t_i' \in T_i^\beta$ while the outcome if t_i follows the strategy from the equilibrium of the auxiliary game is the same in the optimal mechanism given t_i. By incentive compatibility, t_i does not gain by deviating to (s_i, e_i).

So assume (s_i, e_i) is not sent with positive probability by any type in equilibrium. If $\beta = \alpha$ or if there is no $\gamma \neq \alpha$ with $t_i' \in T_i^\gamma$ and $X_i^*(s_i, e_i) = \hat{X}_i^*(t_i')$, then (s_i, e_i) is treated the same way as any $(s_i', e_i') \in R_i^\alpha$ which does have positive probability, so the outcome is the same as if t_i followed the equilibrium strategy from the auxiliary game. Hence, again, he does not gain by deviating.

Finally, suppose (s_i, e_i) has zero probability in equilibrium, $\alpha > \beta$, and there is some $\gamma \neq \alpha$ with $t_i' \in T_i^\gamma$ and $X_i^*(s_i, e_i) = \hat{X}_i^*(t_i')$. If $\gamma > \alpha$, then, again, (s_i, e_i) is treated the same way as any $(s_i', e_i') \in R_i^\alpha$ which has positive probability in the equilibrium of the auxiliary game, so the outcome is again the same as if t_i followed the equilibrium strategy from the auxiliary game. Hence, again, she does not gain by deviating. If $\alpha > \beta > \gamma$, the response to (s_i, e_i) is a 50–50 randomization between the way the principal would respond to positive probability $(s_i', e_i') \in R_i^\alpha$ and the way he would respond to positive probability $(s_i', e_i') \in R_i^\gamma$. This is strictly worse for t_i than the response to t_i's equilibrium strategy. Finally, if $\alpha > \gamma > \beta$, the principal's response is the same as his response to any positive probability $(s_i', e_i') \in R_i^\gamma$, again worse for t_i than following the equilibrium strategy. ▌

D Proof of Theorem 2

We first show there exists v_i^* solving

$$v_i^* = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \text{ or } v_i(t_i) \leq v_i^*]. \tag{4}$$

If $T_i = T_i^0$, then it is easy to see that $v_i^* = \mathrm{E}(v_i(t_i))$ satisfies (4). On the other hand, if $T_i^0 = \emptyset$, then $v_i^* = \min_{t_i \in T_i} v_i(t_i)$ satisfies (4). In what follows, assume $T_i^0 \neq \emptyset$ and $T_i^0 \neq T_i$.

Write $T_i \setminus T_i^0$ as $\{t_i^1, \ldots, t_i^N\}$ where without loss of generality $v_i(t_i^n) < v_i(t_i^{n+1})$. (If we have $t_i, t_i' \in T_i \setminus T_i^0$ with $t_i \neq t_i'$ and $v_i(t_i) = v_i(t_i')$, we can treat these two types as if they were one type for the purposes of this calculation.) For $n = 1, \ldots, N$, let

$$g_i^n = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \text{ or } t_i = t_i^k, \text{ for some } k \leq n]$$

and let $g_i^0 = \mathrm{E}(v_i(t_i) \mid t_i \in T_i^0)$.

Suppose that there is no solution to equation (4). If $g_i^0 \leq v_i(t_i^1)$, then $v_i^* = g_i^0$ satisfies (4). Hence $g_i^0 > v_i(t_i^1)$. But g_i^1 is a convex combination of $v_i(t_i^1)$ and g_i^0, with strictly positive weight on each term, so $v_i(t_i^1) < g_i^1 < g_i^0$. Again, if $g_i^1 \leq v_i(t_i^2)$, then $v_i^* = g_i^1$ satisfies (4), so we must have $g_i^1 > v_i(t_i^2)$, implying $v_i(t_i^2) < g_i^2 < g_i^1$. Similar reasoning gives $g_i^{n-1} > g_i^n > v_i(t_i^n)$ for $n = 1, \ldots, N$. In particular, $g_i^N > v_i(t_i^N)$. But $g_i^N = \mathrm{E}[v_i(t_i)]$, so this implies $v_i^* = g_i^N$ solves equation (4), a contradiction. Hence a solution exists.

To see that the solution is unique, suppose to the contrary that v_i^1 and v_i^2 both solve (4) where $v_i^1 > v_i^2$. Let

$$T_i^k = T_i^0 \ \cup \ \{t_i \in T_i \setminus T_i^0 \mid v_i(t_i) \leq v_i^k\},$$

so $v_i^k = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^k]$. Since $v_i^1 > v_i^2$, we have $T_i^2 \subset T_i^1$ and

$$T_i^1 \setminus T_i^2 = \{t_i \in T_i \setminus T_i^0 \mid v_i^2 < v_i(t_i) \leq v_i^1\}.$$

Note that v_i^1 is a convex combination of v_i^2 and $\mathrm{E}[v_i(t_i) \mid t_i \in T_i^1 \setminus T_i^2]$. But every $t_i \in T_i^1 \setminus T_i^2$ has $v_i(t_i) \leq v_i^1$, so we must have

$$\mathrm{E}[v_i(t_i) \mid t_i \in T_i^1 \setminus T_i^2] \leq v_i^1 \leq v_i^2,$$

contradicting $v_i^1 > v_i^2$.

To construct equilibrium strategies, first note that we must have $x_i^*(s_i, \{t_i\}) = v_i(t_i)$. That is, if t_i proves her type, the principal must infer correctly. Thus we only need to determine the principal's beliefs in response to reports of the form (s_i, T_i) where the agent proves nothing.

It is easy to see that if $T_i^0 = \emptyset$, then $v_i^* = \min_{t_i \in T_i} v_i(t_i)$ and that the essentially unique equilibrium has every type proving her type. This is the usual unraveling argument. Any type t_i' with $v_i(t_i') = \max_{t_i \in T_i} v_i(t_i)$ must strictly prefer proving her type to pooling with lower types and so must prove her type. But then any type with the next highest possible value of $v_i(t_i)$ cannot pool with higher types and so must prove her type, etc. So for the rest of this proof, assume $T_i^0 \neq \emptyset$.

Clearly, we cannot have (s_i, T_i) and (s_i', T_i), both with positive probability in equilibrium with $x_i^*(s_i, T_i) \neq x_i^*(s_i', T_i)$. Since all types are positive, every type strictly prefers whichever of these reports yields the larger x in response. Hence we may as well fix some s_i^* and suppose that the only (s_i, T_i) sent with positive probability in equilibrium is (s_i^*, T_i) where $x_i^*(s_i^*, T_i) \geq x_i^*(s_i, T_i)$ for all $s_i \in T_i$.

Let $\tilde{v}_i = x_i^*(s_i^*, T_i)$. From the above, we know that types $t_i \in T_i^0$ send report (s_i^*, T_i). Any type $t_i \notin T_i^0$ can send either (s_i^*, T_i) and obtain response $\tilde{v}_i$ or can send some $(s_i, \{t_i\})$ and receive response $v_i(t_i)$. Hence t_i chooses the former only if $\tilde{v}_i \geq v_i(t_i)$. Ignoring indifference for a moment, we see that this implies that $\tilde{v}_i$ must be the v_i^* defined in equation (4). To address indifference, note that v^* is not changed if we add or remove from the set of types sending this message a type with $v_i(t_i) = v_i^*$. Hence we have the same outcome regardless. ▌

E Proof of Lemma 1 and Theorem 3

For Lemma 1, the existence and uniqueness of v_i^+ follows from Theorem 2 taking the set of types to be T_i^+. For v_i^-, note that Theorem 2 applied to the function $-v_i(t_i)$ and types T_i^- implies that there is a unique v_i^- satisfying

$$-v_i^- = \mathrm{E}[-v_i(t_i) \mid t_i \in T_i^0 \cap T_i^- \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } -v_i(t_i) \leq -v_i^-)]$$

which can be rewritten as the definition of v_i^-.

Next, we show that there exists v_i^* solving

$$\begin{aligned} v_i^* =& \mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^0) \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) \geq v_i^*) \right. \\ & \left. \text{or } (t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) \leq v_i^*)\right]. \end{aligned} \tag{5}$$

Let the function of v_i^* on the right–hand side be denoted $g_i(v_i^*)$. So we seek to prove that there exists v_i^* solving $v_i^* = g_i(v_i^*)$.

As with Theorem 2, the proof is by contradiction. So suppose there is no v_i^* solving this equation. Let $v_i^1, \ldots, v_i^N$ denote the values of $v_i(t_i)$ for $t_i \notin T_i^0$. Without loss of generality, assume $v_i^k < v_i^{k+1}$ for $k = 1, \ldots, N-1$.

First, note that for $v_i^* \leq v_i^1$, we have $g_i(v_i^*) = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^-]$. So the assumption that there is no v_i^* with $v_i^* = g_i(v_i^*)$ implies $\mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^-] > v_i^1$ as otherwise we can set $v_i^* = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^-]$ and obtain a solution to equation (5).

Clearly, the function $g_i(v_i^*)$ is constant in v_i^* for $v_i^* \in (v_i^k, v_i^{k+1})$. Hence if $g_i(v_i^k) \in (v_i^k, v_i^{k+1})$, we have a solution to equation (5) in this interval.

Also, if $g_i(v_i^k) > v_i^{k+1}$, then $g_i(v_i^{k+1}) > v_i^{k+1}$. To see this, first suppose that $v_i^{k+1} \in v_i(T_i^-)$. In this case, $g_i(v_i^k)$ is a convex combination of $g_i(v_i^{k+1})$ and v_i^{k+1}. Since $g_i(v_i^k) > v_i^{k+1}$ by assumption, we must have $g_i(v_i^{k+1}) \geq g_i(v_i^k) > v_i^{k+1}$, implying the claim. Alternatively, suppose $v_i^{k+1} \in v_i(T_i^+)$. In this case, $g_i(v_i^{k+1})$ is a convex combination of $g_i(v_i^k)$ and v_i^{k+1}. Since $g_i(v_i^k) > v_i^{k+1}$, this implies $g_i(v_i^{k+1}) > v_i^{k+1}$.

As shown above, we start with $g_i(v_i^1) > v_i^1$, so by induction, we have $g_i(v_i^N) > v_i^N$. But $g_i(v_i^*) = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^+]$ for all $v_i^* \geq v_i^N$. So there exists $v_i^* > v_i^N$ solving (5), a contradiction.

To show uniqueness, suppose to the contrary that v_i^1 and v_i^2 are both solutions to equation (5) where $v_i^1 > v_i^2$. Let

$$T_i^{k+} = \{t_i \in T_i^+ \setminus T_i^0 \mid v_i(t_i) \leq v_i^k\}, \quad k = 1, 2$$

and

$$T_i^{k-} = \{t_i \in T_i^- \setminus T_i^0 \mid v_i(t_i) \geq v_i^k\}, \quad k = 1, 2.$$

Clearly, since $v_i^1 > v_i^2$, we have $T_i^{2+} \subseteq T_i^{1+}$ and $T_i^{1-} \subseteq T_i^{2-}$. But

$$v_i^k = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^{k+} \cup T_i^{k-}].$$

Let

$$\tilde{v}_i = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^{2+} \cup T_i^{1-}].$$

Then v_i^1 is a convex combination of $\tilde{v}_i$ and $\mathrm{E}[v_i(t_i) \mid t_i \in T_i^{1+} \setminus T_i^{2+}]$, while v_i^2 is a convex combination of $\tilde{v}_i$ and $\mathrm{E}[v_i(t_i) \mid t_i \in T_i^{2-} \setminus T_i^{1-}]$. It is easy to see that

$$v_i^2 \leq \mathrm{E}[v_i(t_i) \mid t_i \in T_i^{1+} \setminus T_i^{2+}] \leq v_i^1$$

since $v_i^2 \leq v_i(t_i) \leq v_i^1$ for all $t_i \in T_i^{1+} \setminus T_i^{2+}$. Similarly,

$$v_i^2 \leq \mathrm{E}[v_i(t_i) \mid t_i \in T_i^{2-} \setminus T_i^{1-}] \leq v_i^1.$$

Since v_i^1 is a convex combination of $\tilde{v}_i$ and something smaller than v_i^1, we must have $\tilde{v}_i \geq v_i^1$. But since v_i^2 is a convex combination of $\tilde{v}_i$ and something larger than v_i^2, we must have $v_i^2 \geq \tilde{v}_i$. Hence

$$v_i^1 \leq \tilde{v}_i \leq v_i^2,$$

contradicting $v_i^1 > v_i^2$.

Turning to Theorem 3, we construct equilibrium strategies as follows. First, note that if $x_i^*(s_i, T_i) > x_i^*(s_i', T_i)$, then no positive type will send report (s_i', T_i) and no negative type will send (s_i, T_i). Hence there are, at most, two distinct values of $x_i^*(s_i, T_i)$ observed on the equilibrium path. Let $\tilde{v}_i^+ = \max_{s_i \in T_i} x_i^*(s_i, T_i)$ and $\tilde{v}_i^- = \min_{s_i \in T_i} x_i^*(s_i, T_i)$. For the moment, assume $\tilde{v}_i^+ > \tilde{v}_i^-$. Then it is easy to see that every positive type $t_i \in T_i^0$ sends a report generating $\tilde{v}_i^+$ as does every positive type $t_i \notin T_i^0$ with $v_i(t_i) \leq \tilde{v}_i^+$. Similarly, every negative type $t_i \in T_i^0$ or not in T_i^0 but with $v_i(t_i) \geq \tilde{v}_i^-$ sends some report generating $\tilde{v}_i^-$. All other types t_i send a report of the form $(s_i, \{t_i\})$. Given this, it is clear that $\tilde{v}_i^+$ must equal v_i^+ and $\tilde{v}_i^-$ must equal v_i^-. This is an equilibrium iff $v_i^+ \geq v_i^-$. Note that if $v_i^+ = v_i^-$, then the expectation of v_i given the set of types sending either report must also be the same value. Thus in this case, we have $v_i^- = v_i^+ = v_i^*$.

Regardless of the relationship between v_i^- and v_i^+, there is also an equilibrium where the principal's beliefs ignore the type report and condition only on the evidence. Letting $\tilde{v}_i$ denote the principal's expected value of v_i condition on the evidence report $e_i = T_i$, we see that positive types with $v_i(t_i) > \tilde{v}_i$ will prove their types as will negative types with $v_i(t_i) < \tilde{v}_i$. Hence $\tilde{v}_i$ must satisfy equation (5), so $\tilde{v}_i = v_i^*$. ▌

F Proof of Corollary 2

When $v_i^+ \leq v_i^-$, there is only one equilibrium in the auxiliary game for i, so the claim follows. When $v_i^+ > v_i^-$, however, there are (essentially) two equilibria. In one, type reports are used to separate positive types from negative types. All positive types with evidence and $v_i(t_i) > v_i^+$ prove their types, as do all negative types with evidence and $v_i(t_i) < v_i^-$. All other positive types send one type report and evidence $e_i = T_i$, while all other negative types send another type report and the same evidence. In what follows, we refer to this equilibrium as the *cheap–talk equilibrium* as it uses the "cheap talk" of type reports to help separate. In the other equilibrium, the principal's beliefs depend only on the evidence presented, so type reports are irrelevant. All positive types with evidence and $v_i(t_i) > v_i^*$ prove their types as do all negative types with evidence and $v_i(t_i) < v_i^*$. All other types report some fixed type report and evidence $e_i = T_i$. We refer to this equilibrium as the *non–talk equilibrium*.[16]

Since there are two equilibria in the auxiliary game for i in this case, we need to determine which strategies for i are used in the equilibrium of the game which has the same outcome as the optimal mechanism. Clearly, if the principal is better off under one set of strategies than the other, then these must be the strategies used since the equilibrium corresponding to the optimal mechanism must be the best possible equilibrium for the principal.

We now show that the principal's payoff is always at least weakly larger in the cheap–talk equilibrium, completing the proof of Corollary 2.

First, we show that $v_i^+ > v_i^-$ implies $v_i^+ \geq v_i^* \geq v_i^-$ with at least one strict inequality. To see this, suppose to the contrary that $v_i^* > v_i^+ > v_i^-$. Define the following sets of types:

$$\begin{aligned}
\hat{T}_i^- &= \{t_i \in T_i^- \mid t_i \in T_i^0 \text{ or } v_i(t_i) \geq v_i^-\} \\
\hat{T}_i^+ &= \{t_i \in T_i^+ \mid t_i \in T_i^0 \text{ or } v_i(t_i) \leq v_i^+\} \\
\hat{T}_i^{*-} &= \{t_i \in T_i^- \mid t_i \in T_i^0 \text{ or } v_i(t_i) \geq v_i^*\} \\
\hat{T}_i^{*+} &= \{t_i \in T_i^+ \mid t_i \in T_i^0 \text{ or } v_i(t_i) \leq v_i^*\}
\end{aligned}$$

In other words, the types in $\hat{T}_i^-$ are the negative types who "pool" together in the cheap–talk equilibrium, while $\hat{T}_i^+$ is the set of positive types who pool together in this equilibrium. Similarly, $\hat{T}_i^{*-}$ and $\hat{T}_i^{*+}$ are, respectively, the set of negative and positive types who all pool together in the non–talk equilibrium. By definition,

$$v_i^- = \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^-]$$

$$v_i^+ = \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^+]$$

$$v_i^* = \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*-} \cup \hat{T}_i^{*+}]$$

Hence v_i^* is a convex combination of $\mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*-}]$ and $\mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*+}]$.

Since $v_i^- < v_i^*$, we see that $\hat{T}_i^{*-} \subseteq \hat{T}_i^-$. Note that if $t_i \in \hat{T}_i^-$ but $t_i \notin \hat{T}_i^{*-}$, then $v_i^- \leq v_i(t_i) < v_i^*$. Hence $v_i^- = \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^-]$ is a convex combination of $\mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*-}]$ and the expectation of $v_i(t_i)$ for a set of types all with $v_i(t_i) > v_i^-$. Hence

$$v_i^* > v_i^- = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^-] > \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*-}].$$

Similarly, $v_i^+ < v_i^*$ implies that $\hat{T}_i^+ \subseteq \hat{T}_i^{*+}$. Since the types in $\hat{T}_i^{*+} \setminus \hat{T}_i^+$ all satisfy $v_i^+ \leq v_i(t_i) < v_i^*$, we see that $\mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*+}]$ is a convex combination of $v_i^+ = \mathrm{E}[v_i(t_i) \mid$

[16] We refer to this as a non–talk equilibrium rather than as a babbling equilibrium since, unlike in the usual babbling equilibria in the literature, the use of evidence does enable some communication.

$t_i \in \hat{T}_i^+] < v_i^*$ and an expectation of $v_i(t_i)$ for a set of types with $v_i(t_i) < v_i^*$. Hence

$$\mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*+}] < v_i^*.$$

But then we have v_i^* is a convex combination of two terms which are strictly smaller than v_i^*, a contradiction. A similar argument rules out the possibility that $v_i^+ > v_i^- > v_i^*$.

Consider the game between the agents and the principal. We know that there is a robust PBE with the same outcome as in the optimal mechanism. We know i's strategy in this equilibrium must either be the one she uses in the cheap–talk equilibrium or the one she uses in the non–talk equilibrium. Fix the strategies of all agents other than i. We know these strategies are defined from the auxiliary games for these agents, independently of which strategy i uses or the principal's response to i. Thus we can simply determine which strategy by i leads to a higher payoff for the principal.

Note that the principal's payoff for a fixed a is linear in his expectation of v_i. Hence his maximized payoff is convex in his expectation of v_i. We now show that the distribution of beliefs for the principal in the cheap–talk equilibrium is a mean–preserving spread of the distribution in the non–talk equilibrium, completing the proof. To be precise, let (σ_i^1, x_i^1) denote the cheap–talk equilibrium strategies and (σ_i^2, x_i^2) the non–talk equilibrium strategies from the auxiliary game for i. For $k = 1, 2$, define probability distributions B^k over $\mathbf{R}$ by

$$B^k(\hat{v}_i) = \rho_i\left(\left\{t_i \in T_i \mid X_i^k(t_i) = \hat{v}_i\right\}\right).$$

(Recall that $X_i^k(t_i) = x_i^k(s_i, e_i)$ for any (s_i, e_i) with $\sigma_i^k(s_i, e_i \mid t_i) > 0$ and that ρ_i is the prior over T_i.) The law of iterated expectations implies

$$\sum_{\hat{v}_i \in \mathrm{supp}(B^k)} \hat{v}_i B^k(\hat{v}_i) = \mathrm{E}[v_i(t_i)], \quad k = 1, 2.$$

Hence the two distributions have the same mean.

Consider any $\hat{v}_i < v_i^-$. Since $v_i^- \leq v_i^*$, for $k = 1$ or $k = 2$, we have $X_i^k(t_i) = \hat{v}_i$ if and only if there is a negative type with evidence who has $v_i(t_i) = \hat{v}_i$. Similarly, since $v_i^* \leq v_i^+$, for any $\hat{v}_i > v_i^+$, we have $X_i^k(t_i) = \hat{v}_i$ iff there is a positive type with evidence who has $v_i(t_i) = \hat{v}_i$. Hence $B^1(\hat{v}_i) = B^2(\hat{v}_i)$ for any $\hat{v}_i \notin [v_i^-, v_i^+]$.

Also, we have $B^1(\hat{v}_i) = 0$ for all $\hat{v}_i \in (v_i^-, v_i^+)$. Any type with v_i in this range either (1) is positive and chooses to induce belief v_i^+ or (2) is negative and chooses to induce belief v_i^-. Under B^2, however, many of the types generating beliefs concentrated at v_i^- or v_i^+ in the cheap–talk equilibrium instead generate beliefs in (v_i^-, v_i^+). In particular, types without evidence or types with evidence they prefer not to show induce the belief v_i^*, a positive type with evidence who has $v_i(t_i) \in (v_i^*, v_i^+)$ generate the belief $v_i(t_i)$, and similarly for negative types. Hence B^1 is a mean–preserving spread of B^2. ▌

G Costly Verification

In this section, we show that for a particular class of costly verification models with simple type dependence, the optimal mechanism can be computed using our results for optimal mechanisms with Dye evidence. To be specific, we continue to let A denote the finite set of actions available to the principal, T_i the finite set of types of agent i with the same distributional assumptions as in the text, and continue to assume that agent i's utility function can be written as

$$u_i(a, t_i) = \begin{cases} u_i(a), & \text{if } t_i \in T_i^+; \\ -u_i(a), & \text{if } t_i \in T_i^-, \end{cases}$$

and that the principal's utility function can be written as

$$v(a, t) = \sum_{i=0}^{I} u_i(a) v_i(t_i).$$

We add three further assumptions on preferences. First, we assume that each agent has exactly two indifference curves in A.[17] That is, for each agent i, we can partition A into nonempty[18] sets A_i^0 and A_i^1 where

$$u_i(a) = \begin{cases} 0, & \text{if } a \in A_i^0; \\ 1, & \text{if } a \in A_i^1. \end{cases}$$

For example, this assumption holds in the allocation example and most of the related problems discussed in Example 1 of Section 1 as well as the public goods problem discussed in Example 2. It also holds in the public goods problem discussed in Erlanson and Kleiner (2015) (after renormalizing their statement of the payoffs).

Second, we assume that for all i, either $T_i^- = \emptyset$ or $v_i(t_i) > v_i(t_i')$ for all $t_i \in T_i^+$ and $t_i' \in T_i^-$. In other words, either we have type–independent preferences or every positive type has a higher v_i than every negative type. This assumpion on the comparison of positive and negative types is made by Erlanson and Kleiner.

For the costly verification model, the agents do not have evidence to present. Instead, the principal can *check* agent i at a cost $c_i > 0$. "Checking" agent i means that the principal learns agent i's type t_i. We will show that the optimal mechanism for this problem can be computed by an appropriate "translation" of a related mechanism design problem with Dye evidence instead of costly verification.

[17] This also includes "agent 0" — that is, this also applies to the utility function $u_0(a)$. Alternatively, we can simply assume that u_0 is identically zero.

[18] If either set is empty for $i \neq 0$, then the agent is indifferent over all choices by the principal and incentive compatibility is trivially satisfied. Hence we can disregard any such agent.

Note that our assumptions imply that if $v_i(t_i) = v_i(t'_i)$, then either both are positive types or both are negative. Since agents do not have evidence, this means that t_i and t'_i are identical and there is no need for the model (or the principal) to distinguish them. Hence without loss of generality, we assume that if $t_i \neq t'_i$, then $v_i(t_i) \neq v_i(t'_i)$. Consequently, we can write the type set for i as $T_i = \{t_i^0, \ldots, t_i^{K_i}\}$ where $v_i(t_i^k) < v_i(t_i^{k+1})$ for $k = 0, \ldots, K_i - 1$.

It is not hard to prove a simple analog of the Revelation Principle for this class of problems, namely, that it is without loss of generality to focus on mechanisms with the following structure. First, all agents simultaneously make cheap talk reports of types to the principal. The mechanism specifies a probability distribution over which agents to check and what $a \in A$ to choose as a function of the reports. Each agent will have an incentive to report his type honestly, so when the principal checks an agent, he finds that the report was truthful. Off the equilibrium path, if the principal finds that an agent has lied, the principal chooses any action which is worst for that agent. (Since the agents all expect the other agents to report honestly, the specification of the mechanism for histories where multiple agents are found to have lied is irrelevant.)

Hence we can write a mechanism as a function $P : T \to \Delta(2^{\mathcal{I}} \times A)$ where $P(Q, a \mid t)$ is the probability that the principal checks the agents in the set $Q \subseteq \mathcal{I}$ and chooses action $a \in A$ when the type reports are t and the result of the checking verifies that the reports were honest. The expected payoff of the principal from such a mechanism is

$$\mathrm{E}_t \left[\sum_{(Q,a) \in 2^{\mathcal{I}} \times A} P(Q, a \mid t) \left(v(a, t) - \sum_{i \in Q} c_i \right) \right].$$

Let

$$p(a \mid t) = \sum_{Q \subseteq \mathcal{I}} P(Q, a \mid t)$$

$$q_i(t) = \sum_{a \in A} \sum_{Q \subseteq \mathcal{I} \mid i \in Q} P(Q, a \mid t).$$

Then we can rewrite the principal's expected payoff as

$$\mathrm{E}_t \left[\sum_{a \in A} p(a \mid t) v(a, t) - \sum_i q_i(t) c_i \right].$$

Using the fact that $v(a, t) = \sum_i u_i(a) v_i(t_i)$, we can rewrite this as

$$\mathrm{E}_t \left[\sum_i v_i(t_i) \sum_{a \in A} p(a \mid t) u_i(a) - \sum_i q_i(t) c_i \right].$$

Let

$$p_i(t) = \sum_{a \in A_i^1} p(a \mid t).$$

In other words, $p_i(t)$ is the probability that the principal selects an action a such that $u_i(a) = 1$ given type profile t. Then the principal's expected payoff is

$$\mathrm{E}_t \left[\sum_i (p_i(t) v_i(t_i) - q_i(t) c_i) \right] = \sum_i \mathrm{E}_{t_i} [\hat{p}_i(t_i) v_i(t_i) - \hat{q}_i(t_i) c_i],$$

where

$$\hat{p}_i(t_i) = \mathrm{E}_{t_{-i}} p_i(t)$$

and

$$\hat{q}_i(t_i) = \mathrm{E}_{t_{-i}} q_i(t_i, t_{-i}).$$

If agent i of type t_i reports truthfully, his expected utility in mechanism P is

$$\mathrm{E}_{t_{-i}} \sum_{(Q,a) \in 2^{\mathcal{I}} \times A} P(Q, a \mid t) u_i(a)$$

if $t_i \in T_i^+$ and this times -1 otherwise. So the expected payoff to a positive type from reporting truthfully is $\hat{p}_i(t_i)$, while the expected payoff to a negative type is $-\hat{p}_i(t_i)$.

If agent i is type t_i but reports $t_i' \neq t_i$, then he may be caught lying. In this case, as noted above, the principal will choose an action which minimizes his payoff. So if $t_i \in T_i^+$, his payoff will be 0 if he is caught lying, while if $t_i \in T_i^-$, it will be -1. Hence for a positive type, the expected payoff to the deviation is

$$\begin{aligned} &\mathrm{E}_{t_{-i}} \left[\sum_{(Q,a) \in 2^{\mathcal{I}} \times A \mid i \notin Q} P(Q, a \mid t_i', t_{-i}) u_i(a) \right] \\ &\quad = \mathrm{E}_{t_{-i}} \left[\sum_{(Q,a) \in 2^{\mathcal{I}} \times A} P(Q, a \mid t_i', t_{-i}) u_i(a) - \sum_{(Q,a) \in 2^{\mathcal{I}} \times A \mid i \in Q} P(Q, a \mid t_i', t_{-i}) u_i(a) \right] \\ &\quad = \hat{p}_i(t_i') - \mathrm{E}_{t_{-i}} \left[\sum_{(Q,a) \in 2^{\mathcal{I}} \times A \mid i \in Q, a \in A_i^1} P(Q, a \mid t_i', t_{-i}) \right]. \end{aligned}$$

We will simplify this expression further below.

If a negative type is caught reporting falsely, the principal chooses an action setting $u_i(a) = 1$ so that the agent's payoff is -1. Hence the expected payoff to a negative type

t_i from claiming to be $t'_i \neq t_i$ is

$$\begin{aligned}
&\mathrm{E}_{t_{-i}}\left[\sum_{(Q,a)\in 2^{\mathcal{I}}\times A \mid i\notin Q} P(Q,a \mid t'_i, t_{-i})(-u_i(a)) - \sum_{(Q,a)\in 2^{\mathcal{I}}\times A \mid i\in Q} P(Q,a \mid t'_i, t_{-i})\right] \\
&\quad = \mathrm{E}_{t_{-i}}\left[-\sum_{(Q,a)\in 2^{\mathcal{I}}\times A} P(Q,a \mid t'_i, t_{-i})u_i(a) - \sum_{(Q,a)\in 2^{\mathcal{I}}\times A \mid i\in Q} P(Q,a \mid t'_i, t_{-i})(1-u_i(a))\right] \\
&\quad = -\hat{p}_i(t'_i) - \mathrm{E}_{t_{-i}}\left[\sum_{(Q,a)\in 2^{\mathcal{I}}\times A \mid i\in Q, a\in A_i^0} P(Q,a \mid t'_i, t_{-i})\right].
\end{aligned}$$

Summarizing, the incentive compatibility constraint for agent i is that for all positive types $t_i \in T_i^+$, we have

$$\hat{p}_i(t_i) \geq \hat{p}_i(t'_i) - \mathrm{E}_{t_{-i}}\left[\sum_{(Q,a)\in 2^{\mathcal{I}}\times A \mid i\in Q, a\in A_i^1} P(Q,a \mid t'_i, t_{-i})\right], \ \forall t'_i \neq t_i \tag{6}$$

and for all negative types $t_i \in T_i^-$, we have

$$\hat{p}_i(t_i) \leq \hat{p}_i(t'_i) + \mathrm{E}_{t_{-i}}\left[\sum_{(Q,a)\in 2^{\mathcal{I}}\times A \mid i\in Q, a\in A_i^0} P(Q,a \mid t'_i, t_{-i})\right], \ \forall t'_i \neq t_i. \tag{7}$$

Note that the right–hand side of each incentive compatibility constraint is independent of t_i. Hence (6) holds for all positive types t_i iff it holds for the positive type with the smallest $\hat{p}_i(t_i)$ and (7) holds for all negative types t_i iff it holds for that negative type with the largest $\hat{p}_i(t_i)$.

It is not hard to show that the optimal mechanism must be monotonic in the sense that $\hat{p}_i(t_i^k) \leq \hat{p}_i(t_i^{k+1})$ for $k = 0, \ldots, K_i - 1$. To see this, recall that $v_i(t_i^k) < v_i(t_i^{k+1})$, so the principal is better off with higher values of p_i associated with higher values of t_i. So suppose we have an incentive compatible mechanism with $\hat{p}_i(t_i^k) > \hat{p}_i(t_i^{k+1})$ for some k and i. Consider the mechanism which reverses the roles of these types — i.e., assigns the outcome (Q,a) to (t_i^k, t_{-i}) that it would have assigned to (t_i^{k+1}, t_{-i}) and vice versa.[19] Then this altered mechanism is also incentive compatible and yields the principal a higher expected payoff.

By assumption, for every i, either $T_i^- = \emptyset$ or $v_i(t_i) > v_i(t'_i)$ for all $t_i \in T_i^+$, $t'_i \in T_i^-$. Hence if there are J_i negative types(where J_i can be zero), the negative types are $t_i^0, \ldots, t_i^{J_i-1}$ and the positive types are $t_i^{J_i}, \ldots, t_i^{K_i}$. Thus the positive type with the

[19] To be precise, this implicitly assumes the two types have the same prior probability. If not, we can reverse the role of one of the types and "part of" the other.

lowest $\hat{p}_i(t_i)$ is $t_i^{J_i}$, while the negative type with the highest $\hat{p}_i(t_i)$ is $t_i^{J_i-1}$ and we have $\hat{p}_i(t_i^{J_1-1}) \leq \hat{p}_i(t_i^{J_i})$. So we can write the incentive compatibility constraints (6) and (7) as

$$\hat{p}_i(t_i^{J_i}) \geq \hat{p}_i(t_i') - \mathrm{E}_{t_{-i}}\left[\sum_{(Q,a)\in 2^{\mathcal{I}}\times A \mid i\in Q, a\in A_i^1} P(Q,a \mid t_i', t_{-i})\right], \ \forall t_i' \neq t_i \tag{8}$$

and

$$\hat{p}_i(t_i^{J_i-1}) \leq \hat{p}_i(t_i') + \mathrm{E}_{t_{-i}}\left[\sum_{(Q,a)\in 2^{\mathcal{I}}\times A \mid i\in Q, a\in A_i^0} P(Q,a \mid t_i', t_{-i})\right], \ \forall t_i' \neq t_i. \tag{9}$$

The following lemma simplifies the incentive compatibility constraints.

Lemma 13. *In any optimal mechanism, we have*

$$P(Q,a \mid t_i, t_{-i}) = 0, \ \forall t_{-i} \text{ if } t_i \in T_i^+, \ i \in Q, \text{ and } a \in A_i^0$$

and

$$P(Q,a \mid t_i, t_{-i}) = 0, \ \forall t_{-i} \text{ if } t_i \in T_i^-, \ i \in Q, \text{ and } a \in A_i^1.$$

Consequently, we can rewrite the incentive compatibility constraints (8) and (9) as

$$\hat{p}_i(t_i^{J_i}) \geq \hat{p}_i(t_i) - \hat{q}_i(t_i), \ \forall t_i \in T_i^+ \tag{10}$$

and

$$\hat{p}_i(t_i^{J_i-1}) \leq \hat{p}_i(t_i) + \hat{q}_i(t_i), \ \forall t_i \in T_i^-. \tag{11}$$

Proof. First, we show that we only require (8) for $t_i' \in T_i^+$ and (9) for $t_i' \in T_i^-$. Specifically, we show that monotonicity of $\hat{p}_i$ implies that (8) holds for all $t_i' \in T_i^-$ and (9) holds for all $t_i' \in T_i^+$. To see this, fix any $t_i' \in T_i^-$. By assumption, $v_i(t_i') \leq v_i(t_i^{J_i})$, so monotonicity implies $\hat{p}_i(t_i^{J_i}) \geq \hat{p}_i(t_i')$. Since $\hat{p}_i(t_i')$ is weakly larger than the right–hand side of (8), this implies (8) holds. A similar argument shows that monotonicity implies (9) for any $t_i' \in T_i^+$.

Next, suppose, contrary to the statement of the lemma, that we have an optimal mechanism P with the property that $P(Q,a \mid t_i, t_{-i}) > 0$ for some $t_{-i} \in T_{-i}$, $t_i \in T_i^+$, $i \in Q$, and $a \in A_i^0$. In other words, there is a positive probability that the principal checks some positive type and then chooses an action giving that agent a payoff of zero. Construct a new mechanism P^* as follows. For any $(Q', a') \neq (Q, a)$ or $t' \neq t$, let $P^*(Q', a' \mid t') = P(Q', a' \mid t')$. Let $P^*(Q, a \mid t) = 0$ and let $P^*(Q \setminus \{i\}, a \mid t) = P(Q, a \mid t) + P(Q \setminus \{i\}, a \mid t)$. In other words, if i is checked but gets a zero payoff at (Q, a), we shift this probability to $(Q \setminus \{i\}, a)$, where i does not get checked but still gets the same zero payoff. It is easy to see that the incentive compatibility constraints for any agent $j \neq i$ are unaffected. Since t_i is a positive type, the only incentive compatibility constraint for i that is potentially affected is (8) at $t'_i = t_i$ or where $t_i = t_i^{J_i}$. But since we have only changed the checking probability and not the marginal probabilities over actions $a \in A$, $\hat{p}_i(t_i)$ is unaffected by this change in the mechanism. Similarly, the second term on the right–hand side of (8) for $t'_i = t_i$ only involves actions in A_i^1, so this term also is unaffected. Hence P^* is incentive compatible. Finally, since the probability over A is unchanged but the principal checks less often, his payoff must be strictly larger, a contradiction.

So suppose we have an optimal mechanism P with the property that $P(Q, a \mid t_i, t_{-i}) > 0$ for some $t_{-i} \in T_{-i}$, $t_i \in T_i^-$, $i \in Q$, and $a \in A_i^1$, contrary to the statement of the lemma. That is, we have a strictly positive probability that the principal checks some negative type and then chooses an action giving that agent a payoff of -1. Construct a new mechanism P^* exactly as in the previous case. That is, for any $(Q', a') \neq (Q, a)$ or $t' \neq t$, we do not change the mechanism, so $P^*(Q', a' \mid t') = P(Q', a' \mid t')$. Again, we let $P^*(Q, a \mid t) = 0$ and let $P^*(Q \setminus \{i\}, a \mid t) = P(Q, a \mid t) + P(Q \setminus \{i\}, a \mid t)$. Again, it is easy to see that the incentive compatibility constraints for any agent $j \neq i$ are unaffected. Since t_i is a negative type, the only incentive compatibility constraint for i that is potentially affected is (9) at $t'_i = t_i$ or where $t_i = t_i^{J_i - 1}$. But since we have only changed the checking probability and not the probabilities over actions, $\hat{p}_i(t_i)$ is unaffected by this change in the mechanism. Analogously to the previous case, the second term on the right–hand side of (9) only involves actions in A_i^0, so this term also is unaffected. Hence P^* is incentive compatible. Finally, just as before, the probability over A is unchanged but the checking probabilities are lower, making the principal strictly better off, a contradiction.

To conclude, consider equation (8) for t'_i in light of the above. Since $P(Q, a \mid t'_i, t_{-i}) = 0$ if $a \in A_i^0$, we see that

$$\sum_{(Q,a) \in 2^{\mathcal{I}} \times A \mid i \in Q, a \in A_i^1} P(Q, a \mid t'_i, t_{-i}) = \sum_{(Q,a) \in 2^{\mathcal{I}} \times A \mid i \in Q} P(Q, a \mid t'_i, t_{-i}) = q_i(t'_i, t_{-i}).$$

Hence we can rewrite (8) as $\hat{p}_i(t_i^{J_i}) \geq \hat{p}_i(t'_i) - \hat{q}_i(t'_i)$ for all $t'_i \in T_i^+$. A similar argument applied to (9) completes the proof. ▌

In light of Lemma 13, we can directly compute $\hat{q}_i(t'_i)$ for all t'_i. Since $\hat{q}_i$ is costly for the principal, we see that the inequalities in equations (10) and (11) must hold with equality, so

$$\hat{q}_i(t_i) = \begin{cases} \hat{p}_i(t_i) - \hat{p}_i(t_i^{J_i}), & \text{if } t_i \in T_i^+; \\ \hat{p}_i(t_i^{J_i - 1}) - \hat{p}_i(t_i), & \text{if } t_i \in T_i^-. \end{cases}$$

We can use this to substitute into the objective function for $\hat{q}_i$ to rewrite it as

$$\sum_i \mathrm{E}_{t_i}[\hat{p}_i(t_i)v_i(t_i) - \hat{q}_i(t_i)c_i] = \sum_i \Bigg[\sum_{k=0}^{J_i - 1} \rho_i(t_i^k)[\hat{p}_i(t_i^k)(v_i(t_i^k) + c_i) - \hat{p}_i(t_i^{J_i - 1})c_i] \tag{12}$$
$$+ \sum_{k=J_i}^{K_i} \rho_i(t_i^k)[\hat{p}_i(t_i^k)(v_i(t_i^k) - c_i) + \hat{p}_i(t_i^{J_i})c_i] \Bigg].$$

The only remaining incentive constraints are that $\hat{p}_i(t_i) \leq \hat{p}_i(t_i^{J_i - 1}) \leq \hat{p}_i(t_i^{J_i})$ for all negative types t_i and $\hat{p}_i(t_i^{J_i - 1}) \leq \hat{p}_i(t_i^{J_i}) \leq \hat{p}_i(t_i)$ for all positive types t_i.

Now consider a different mechanism design problem, this one with evidence instead of costly verification. We have the same set of types as in the problem above and the same u_i functions. As above, types $t_i^0, \ldots, t_i^{J_i - 1}$ are negative and types $t_i^{J_i}, \ldots, t_i^{K_i}$ are positive. The principal's objective function is now

$$\sum_i u_i(a)\tilde{v}_i(t_i)$$

where

$$\tilde{v}_i(t_i) = \begin{cases} v_i(t_i) - c_i, & \text{if } t_i \in T_i^+ \text{ and } t_i \neq t_i^{J_i}; \\ v_i(t_i) + c_i, & \text{if } t_i \in T_i^- \text{ and } t_i \neq t_i^{J_i - 1}; \\ v_i(t_i^{J_i}) - c_i + \frac{c_i}{\rho_i(t_i^{J_i})}, & \text{if } t_i = t_i^{J_i}; \\ v_i(t_i^{J_i - 1}) + c_i - \frac{c_i}{\rho_i(t_i^{J_i - 1})}, & \text{if } t_i = t_i^{J_i - 1}. \end{cases}$$

It is easy to see that this specification of $\tilde{v}_i$ makes the principal's objective function in this problem the same as the expression in equation (12).

We specify the evidence structure as follows. For any t_i other than $t_i^{J_i - 1}$ or $t_i^{J_i}$, we have $\mathcal{E}_i(t_i) = \{\{t_i\}, T_i\}$. Also, $\mathcal{E}_i(t_i^{J_i - 1}) = \mathcal{E}_i(t_i^{J_i}) = \{T_i\}$. The incentive compatibility constraints for this problem, then, are the following. First, since types $t_i^{J_i - 1}$ and $t_i^{J_i}$ can each claim to be the other and send the other's (trivial) maximal evidence, each must weakly prefer her own allocation. Since $t_i^{J_i - 1}$ is a negative type and $t_i^{J_i}$ is positive, this implies $\hat{p}_i(t_i^{J_i - 1}) \leq \hat{p}_i(t_i^{J_i})$. This implies that any other negative type prefers imitating $t_i^{J_i - 1}$ to imitating $t_i^{J_i}$, while any positive type has the opposite preference. Hence the only other incentive compatibility constraints are $\hat{p}_i(t_i) \leq \hat{p}_i(t_i^{J_i - 1})$ for any negative type t_i and $\hat{p}_i(t_i) \geq \hat{p}_i(t_i^{J_i})$ for any positive type t_i, exactly the same constraints as in the costly verification model.

Hence we can apply our results on optimal mechanisms with Dye evidence to compute the optimal mechanism for the evidence model as a function of $\tilde{v}_i$. We can then substitute in terms of v_i to rewrite in terms of the original costly verification model. It is straightforward to show that doing so for the case considered in Ben-Porath, Dekel, and Lipman (2014) or for the case considered in Erlanson and Kleiner (2015) yields the optimal mechanism identified there.

Because the assumptions used here also cover several of the variations of Example 1 discussed in Section 1, we can also use this approach and the characterization given in Examples 4 and 5 of Section 3.1 to characterize optimal mechanisms with costly verification for the case where the principal allocates multiple identical goods or the case where he allocates a "bad."

While this connection between costly verification mechanisms and Dye evidence mechanisms is likely to hold for some more general assumptions than we have used, some assumption beyond simple type dependence is necessary for it. For example, certain versions of the task allocation problem which satisfy simple type dependence but where $v_i(t_i) > v_i(t_i')$ for some $t_i \in T_i^-$ and $t_i' \in T_i^+$ for some i do not satisfy this equivalence property.

References

[1] Ben-Porath, E., E. Dekel, and B. Lipman, “Optimal Allocation with Costly Verification,” *American Economic Review*, **104**, December 2014, 3779–3813.

[2] Ben-Porath, E., and B. Lipman, “Implementation and Partial Provability,” *Journal of Economic Theory*, **147**, September 2012, 1689–1724.

[3] Bull, J., and J. Watson, “Hard Evidence and Mechanism Design,” *Games and Economic Behavior*, **58**, January 2007, 75–93.

[4] Deneckere, R. and S. Severinov, “Mechanism Design with Partial State Verifiability,” *Games and Economic Behavior*, **64**, November 2008, 487–513.

[5] Dye, R. A., “Disclosure of Nonproprietary Information,” *Journal of Accounting Research*, **23**, 1985, 123–145.

[6] Erlanson, A., and A. Kleiner, “Costly Verification in Collective Decisions,” working paper, November 2015.

[7] Fudenberg, D., and J. Tirole, “Perfect Bayesian Equilibrium and Sequential Equilibrium,” *Journal of Economic Theory*, **53**, April 1991, 236–260.

[8] Glazer, J., and A. Rubinstein, “On Optimal Rules of Persuasion,” *Econometrica*, **72**, November 2004, 1715–1736.

[9] Glazer, J., and A. Rubinstein, “A Study in the Pragmatics of Persuasion: A Game Theoretical Approach,” *Theoretical Economics*, **1**, December 2006, 395–410.

[10] Green, J., and J.-J. Laffont, “Partially Verifiable Information and Mechanism Design,” *Review of Economic Studies*, **53**, July 1986, 447–456.

[11] Grossman, S. J., “The Informational Role of Warranties and Private Disclosures about Product Quality,” *Journal of Law and Economics*, **24**, 1981, 461–483.

[12] Hart, S., I. Kremer, and M. Perry, “Evidence Games: Truth and Commitment,” *American Economic Review*, forthcoming.

[13] Hart, S., I. Kremer, and M. Perry, “Evidence Games with Randomized Rewards,” working paper in progress, 2016.

[14] Jung, W., and Y. Kwon, “Disclosure When the Market is Unsure of Information Endowment of Managers,” *Journal of Accounting Research*, **26**, 1988, 146–153.

[15] Kartik, N., and O. Tercieux, “Implementation with Evidence,” *Theoretical Economics*, **7**, May 2012, 323–355.

[16] Lipman, B., and D. Seppi, “Robust Inference in Communication Games with Partial Provability,” *Journal of Economic Theory*, **66**, August 1995, 370–405.

[17] Milgrom, P., “Good News and Bad News: Representation Theorems and Applications,” *Bell Journal of Economics*, **12**, 1981, 350–391.

[18] Sher, I., “Credibility and Determinism in a Game of Persuasion,” *Games and Economic Behavior*, **71**, March 2011, 409–419.

[19] Sher, I., and R. Vohra, “Price Discrimination through Communication,” *Theoretical Economics*, **10**, May 2015, 597–648.